AD_____

Award Number: DAMD17-97-1-7130

TITLE: Computer-Assisted Visual Search/Decision Aids as a Training Tool for Mammography

PRINCIPAL INVESTIGATOR: Calvin Nodine, Ph.D.

CONTRACTING ORGANIZATION: University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

REPORT DATE: July 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2000	Annual (1 Jul 99 - 30 Jun 00)

4. TITLE AND SUBTITLE
Computer-Assisted Visual Search/Decision Aids as a Training Tool for Mammography

5. FUNDING NUMBERS
DAMD17-97-1-7130

6. AUTHOR(S)
Calvin Nodine, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

E*Mail: nodine@oasis.rad.upenn.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The primary goal of the project is to develop a computer-assisted visual search (CAVS) mammography training tool that will improve the perceptual and cognitive skills of trainees leading to mammographic expertise. In the first two years we carried out two experiments. The first equated experience by comparing perceptual skills of expert radiologists with laypeople searching non-medical pictorial scenes for hidden targets. Results show that expert radiology search and detection strategies do not transfer to the non-medical search and detection tasks. In the second study, a 75-case mammogram test set was administered to mammographers, residents and mammography technologists. The results compared effectiveness of experience and training at different levels of expertise. Not surprisingly, resident performance in detecting and classifying breast lesions was significantly inferior to experts, and no better than that of mammography technologists. In the third year we carried out an experiment to determine if retrospectively-visible cancers in mammograms attract visual attention and are accurately recognized in a blinded review. Results comparing a test set of 40 retrospectively-visible vs. directly-visible cancer cases indicated that not only did retrospectively-visible cancers fail to attract visual attention, but that they also led to higher false-positive error rates than did directly-visible cancers.

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
64

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

N/A Where copyrighted material is quoted, permission has been obtained to use such material.

N/A Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

N/A Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

CPT Wodine 7/18/00

PI date

Table of Contents

PROGRESS REPORT, 1999-2000, Year 3, DAMD17-97-1-7130 COMPUTER-ASSISTED VISUAL SEARCH/DECISION AIDS AS A TRAINING TOOL FOR MAMMOGRAPHY.
C.F. NODINE, PI 7/18/2000

(5) INTRODUCTION:

This project focuses on the perceptual training of diagnostic interpretation skills in mammography which are acquired mainly as a result of experience reading mammograms. The primary aim of this project is to develop a computer-assisted mammography training tool that will act as a surrogate mentor in aiding radiologists in making plausible diagnostic decisions. We propose to provide a computer aid that will interact with the radiologist immediately after image interpretation by providing systematic feedback about how the mammogram was searched for abnormalities and what features received prolonged visual attention indicating potential lesions during scanning. The eye-position parameter, visual dwell, is used to predict the locations of suspicious lesions on the mammogram (Krupinski, Nodine, Kundel, 1998; Nodine, Kundel, Mello-Thoms et al., 1999). The resident is then asked to re-examine the highlighted areas, determine if any abnormal features are present, and re-evaluate the original diagnostic decision. This re-evaluation of suspicious regions with visual feedback provides a perceptually-guided basis for a plausible problem-solving diagnostic solution. We showed in 1990 (Kundel, Nodine, Krupinski, 1990) that computer-assisted visual search (CAVS) is effective in improving the detection of lung nodules, and Krupinski (1996) showed that visual dwell predicts the location of true and false, positive and negative decision outcomes. Our goal is to determine if CAVS improves the detection and interpretation of breast cancers.

(6) BODY:

(6.1) OBJECTIVES. The primary objective of the work this year was to develop a working CAVS. This task was delayed because of technical problems interfacing the ASL 4000SU eye-head tracker to the Microsoft WINDOWS 95 environment. We have already completed Tasks 3, 4 and part of 5 as reported in last year's progress report. We will now report on work completed from July 1, 1999 to June 31, 2000 based on the approved Statement of Work.

(6.2) TECHNICAL OBJECTIVE 1, DEVELOP A COMPUTER-ASSISTED VISUAL SEARCH (CAVS) SYSTEM AND DIGITAL DISPLAY WORKSTATION.

(6.3) Task 1. Program ASL Model 4000SU and EYEHEAD to monitor the observer's eye position relative to head motion for digital mammography displays. We have completed TASK 1. We have programmed the ASL Model 4000 to monitor the observer's eye position relative to head motion for digital mammography displays. In addition, we have programmed the workstation to record viewing time, and event times associated with observer localizations and decisions of breast lesions discovered during visual scanning of

the digital mammogram display. **TASK 1 Interface ASL 4000SU System with Display Workstation, COMPLETE.**

(6.4) Task 2. Modify eye-position data collection programs (EYEPOS/EYEDAT) to accommodate visual-dwell detection algorithm. Integrate detection algorithm with visual feedback of dwell locations on PC display workstation. This was a difficult task to complete from a technical standpoint because the ASL 4000SU eye-head tracker was programmed and operated within a DOS environment and the PC workstation operated within a WINDOWS 95 environment. However, we obtained a software DLL driver from ASL and have integrated it into our workstation display successfully. In addition, we have accomplished the necessary programming to record, analyze and store eye-head position data and, after initial observer evaluation, feedback by highlighting mammogram features that receive prolonged dwell. Thus, we now have a working CAVS system for mammography, and are ready to begin the final phase of testing it with mammographers. **TASK 2 Program to Modify, Analyze and Display Eye Position Data, COMPLETE.**

(6.5) Task 5. Carry out pilot study to determine the effectiveness of the integration of the CAVS dwell-detection designed to help differentiate true from false positive and negative decisions in the mammography interpretation task. We have completed a study using the ASL 4000SU to monitor mammographers' eye-head position during mammography interpretation (Nodine, Mello-Thoms, Weinstein et al., 2000, in preparation). The aim of this study was to determine if retrospectively identified cancers in mammograms can be reliably recognized in a blinded review. These retrospectively identified cancers were not reported on initial screening, but were reported subsequently (average screening interval 14 months). The question that arises is whether these cancers were initially missed, or were they so subtle that they were impossible to detect? This question has been addressed before, but never with the benefit of eye position data to determine if such cancers attract visual attention.

Observers were 4 experienced mammographers, who performed a blinded review on a test set of 20 retrospectively visible but unreported (U) cancer cases, 10 reported (R) cancer cases, and 10 cancer-free cases. Two views were digitized and displayed on our high-resolution digital workstation. The study had two phases: Phase 1 Perception, during which eye-position was monitored; Phase 2 Interpretation, during which observers zoomed on regions of interest and localized suspicious lesions. All of these events were automatically recorded by our new CAVS system.

Eye-position data were analyzed to determined if observers fixated the subtle previous unreported cancers (U cases), and to compare this performance to that of previously detected cancers (R cases). Using a 1000 ms visual dwell threshold and an overall decision "Abnormal" as the criteria for a decision event yielded hypothetical case performance for Phase 1 Perception. Significantly more TPs and fewer false positives were fixated for R cases than for U cases. In Phase 2 Interpretation, zooming to magnify suspicious features on the mammogram decreased FPs for both case types, but only increased TPs for R cases. We will discuss the implications of these findings below. We are currently writing

up this study for publication. A final draft is expected by the end of July. Thus, we have tested the CAVS system and are now ready to perform a final test using CAVS with visual feedback. **TASK 5 Pilot Study of CAVS, COMPLETE.**

(6.6) IMPLICATIONS OF PILOT STUDY OF RETROSPECTIVELY VISIBLE BUT UNREPORTED BREAST CANCERS. On blinded review, meaning that observers had no prior knowledge of the test cases, the performance of experienced mammographers on retrospectively visible but unreported breast cancer cases (U cases) was significantly inferior to that of reported breast cancer cases (R cases). Table 1 shows the results.

Table 1

Number of Lesions Identified Either by Eye Fixations in Phase 1, or by Observer Localization and Interpretation in Phase 2 and Overall Performance as Measured by d' for Unreported (U) and Reported (R) Cases.

		Unreported (U) Cases		Reported (R) Cases	
		True Lesions	Non-Lesions	True Lesions	Non-Lesions
Phase 1 Perception:	O1	.45	.20	.60	.30
At Least One Fixation	O2	.15	.30	.70	.40
Cluster > 1000 ms	O3	.10	.10	.10	.01
	O4	.50	.20	.60	.40
Average		.30	.20	.50	.28
		d'= 0.32		d'= 0.58	
Phase 2 Decision Making:	O1	.65	.20	.90	.10
At Least One Lesion	O2	.20	.45	.70	.20
Localized and Interpreted	O3	.30	.15	.70	.01
	O4	.60	.35	.70	.01
Average		.44	.29	.75	.08
		d'= 0.41		d'= 2.08	

Note: Total number for U cases was 20 x 4 observers= 80, and total number for R cases was 10 x 4 observers= 40. TP hits and FP hits in Phase 1 were defined by eye-fixations falling on true lesions or non-lesions rather than actual observer reports. There were 5/80= .06 U true lesions reported Abnormal but not fixated, and 2/40= .05 R true lesions reported Abnormal but not fixated.

Eye position data indicated that most of the U cancers judged visible in retrospect (by an experienced mammographer who did not participate as an observer in the study) did not attract visual attention during the Perceptual Phase. Overall performance for U cases in terms of fixating a true cancer long enough (>1000 ms) to render an "Abnormal" case decision was significantly below chance (Chi Square test = 12.8, $p<.001$). After the Interpretation Phase during which mammographers were able to examine the cases in detail with zooming and roving overall performance as measured by the index of detectability, d', significantly improved for R cases (d'= 2.08, Analysis of Variance, $F(1,6)= 31.22$, $p< .01$), but not for U cases (d'= .41).

Detailed examination of mammograms by zooming led to the discovery of new true cancer cases: 31% for U cases; and, 33% for R cases. But, the increase in the discovery of true cancer U cases was offset by a 30% increase in false positive cases resulting in only a 1% net gain. This suggests that mammographers were operating at close to chance in picking up new cancer cases from detailed examination in Phase 2. For R cases, false positive cases decreased by 73% resulting in a 106% net gain where net gain= (true positives) – (false positives).

Table 2 shows overall performance as measured by area under the AFROC, curve (which stands for Alternative Free Response Operating Characteristic) for U and R cases when the unit of analysis is lesion rather than case. This analysis is more strict because cancers have to be localized and reported by the mammographers. The results in Table 2 are the culmination of the Perceptual and Interpretation Phases.

Table 2

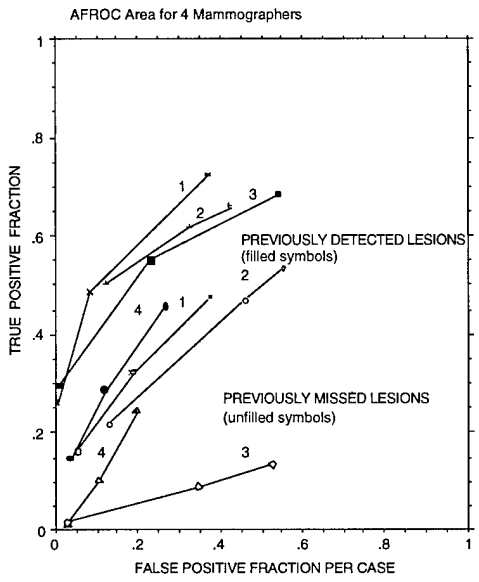
A1 Areas Under AFROC for Unreported (U) and Reported (R) Cases Based on Localization and Interpretation of Lesions in CC and MLO views per Case.

	Unreported Cases, U		Reported Cases, R	
	A1	sd	A1	sd
O1	.554	(.086)	.751	(.080)
O2	.159	(.067)	.656	(.087)
O3	.623	(.129)	.622	(.117)
O4	.497	(.074)	.673	(.090)
Average	.458	(.089)	.675	(.094)

The mean A1 areas, which represent overall performance in correctly interpreting malignant lesions, for U cases was .458 (.089) v. mean A1 areas for R cases which was .676 (.094), and this difference was significant (t-test, (6)= 3.37, $p < .05$). The A1 areas range from +1 to -1. The scoring of performance in this analysis took into account localization of lesions in both CC and MLO images for each case. Thus of the 126 lesions reported as malignant for U cases only 41% (52/126) were correctly localized; for R cases, 104 lesions were reported as malignant and 63% (65/104) were correctly localized.

The expected operating points on AFROC curves are plotted for each observer in Figure 1. These points represent the predicted intersections of true positives (y-axis) and false positives (x-axis) using maximum-likelihood estimates. The 3 estimated operating points for each observer in Figure 1 show how performance, as measured by AFROC area, increases as a function of cumulative decision confidence using a 3-point scale of low, medium and high.

Figure 1. AFROC Estimated Operating Points for 4 Mammographers.

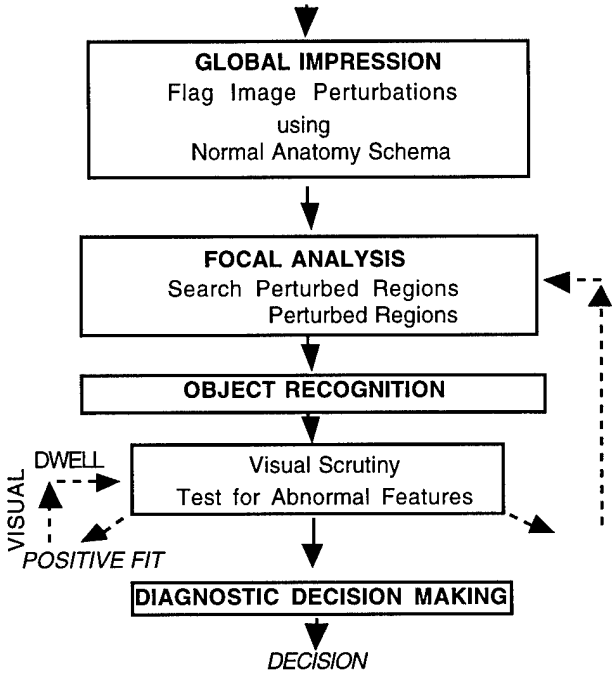


Performance of all 4 mammographers was higher for previously detected lesions (R cases) than for previously missed lesions (U cases), and all but mammographer no. 3 retained their rank order between the two sets of cases. Observers 3 and 4 were operating close to chance at lowest decision confidence for previously missed lesions (U cases). This would be indicated by a 0 true positive fraction on the plot in Figure 1.

The pattern of performance yielding such a small net gain, even from detailed evaluation by zooming for U cases, suggests that most of the true cancer cases were discovered during initial scanning in Phase 1. However, for R cases zooming definitely did benefit performance with a solid net gain in new true cancer cases detected in Phase 2.

Our findings suggest that the Global Impression plays a major role in flagging image perturbations that are recognized as deviating from normal anatomy (Nodine, Mello-Thoms, 2000). A recent model that we have proposed (see Figure 2) shows the relationship between Global Impression and Focal Analysis.

Figure 2 A Perceptual Model of the Radiology Task.



(7) KEY RESEARCH ACCOMPLISHMENTS:

Our research studies in 1999-2000 have led to three key findings:

1. For subtle cancer cases where the cancer was not detected at first screening, the Global Impression that initiates image perception plays a major role in guiding perceptual analysis during visual search and interpretation of mammograms.

2. Detailed examination of digital mammograms with a zoom magnification tool is used by experienced mammographers primarily to confirm initial perceptions flagged during Global Impression rather than to discover new abnormalities.

3. Retrospective analysis of missed cancer cases biases mammographers' judgments and lead to illogical conclusions in deciding whether or not to report a lesion as malignant.

(8) REPORTABLE OUTCOMES:

In addition to the work completed and in progress as discussed above, we have completed the following articles:

1. "How Experience and Training Influence Mammography Expertise", ACADEMIC RADIOLOGY, 1999; 6: 575-585. (see Nodine, Kundel, Mello-Thoms et al., 1999, in press, Appendix 1).

2. "Do Subtle Cancers Attract Visual Attention During Initial Impression?" was presented at Medical Imaging 2000, SPIE PROCEEDINGS 2000; 3981:156-159 (see Nodine, Mello-Thoms, Kundel, et al. 2000, Appendix 2).

3. "A Perceptually Tempered Display for Digital Mammograms", RADIOGRAPHICS, 1999; 19: 1313- 1318 (see Kundel, Weinstein, Conant, Toto, Nodine, 1999, in press, Appendix 3).

4. " Image Structure and Perceptual Errors in Mammogram Reading: A Pilot Study" was presented at Medical Imaging 2000, SPIE PROCEEDINGS 2000; 3981:170-173. (see Mello-Thoms, Dunn, Nodine et al., 2000, Appendix 4).

5. "An Unobtrusive Method for Monitoring Visual Attention During Mammogram Reading" was presented at Medical Imaging 2000, SPIE PROCEEDINGS 2000; 3981:160-163. (see Mello-Thoms, Nodine, Weinstein et al., 2000, Appendix 5).

6. "The Nature of Expertise in Radiology", Chapter 19 in the Handbook of Medical Imaging, Volume 1. Physics and Psychophysics, Edited by J. Beutel, H.L. Kundel, R. L. Van Metter Bellingham, WA: SPIE Press, 2000; 859-894. (see Nodine, Mello-Thoms, 2000, Appendix 6)

We are currently working on three papers:

1. "A Model for Information Acquition in Reading Medical Images: Chest Radiographs vs. Mammograms" Mello-Thoms, Dunn, Nodine et al, 2000, in preparation.

2. "Blinded Review of Retrospectively Visible But Unreported Breast Cancers: An Eye-Position Analysis" Nodine, Mello-Thoms, Weinstein et al., 2000, in preparation.

3. "An analysis of perceptual errors in reading mammograms using quasi-local image frequency spectra. Mello-Thoms, Dunn, Nodine CF et al., 2000, in preparation.

(9) CONCLUSIONS

The primary goal of the project is to develop a mammography training tool that will improve perceptual and cognitive skills of observers leading to mammographic expertise.

Prerequisites to this goal are an understanding of: (a) how mammographers are trained, (b) what skills are required to carry out the task of detecting, classifying and diagnosing abnormalities in mammograms, and (c) the effectiveness of current mammography training measured by evaluating the performance of residents using a test-set of mammograms representing various abnormalities. We have examined these three questions and reported the results in two articles (Nodine, Kundel, Mello-Thoms, 1999; Nodine, Mello-Thoms, 2000).

We have shown that the amount of experience reading mammogram cases with a mentor (defined as deliberate practice) has significant impact on overall diagnostic performance. The residents that we studied at the University of Pennsylvania received an average of 645 case-reading experiences which from our regression analysis leads to a performance prediction that is well below acceptable clinical standards. This brings us to the question of what skills need to be improved, and how can this be accomplished.

Our research has focused on perceptual and decision-making skills in mammography. We have used eye-position recording to shed light on the role of visual search in diagnostic performance. Visual search skills translate into rapid image-perception assessment which leads to fast, accurate decision making as indicated by decision-time analyses. We have called this the speed-accuracy relationship.

Finally, when we come to the question of how can perceptual and decision-making skills be improved? The answer that our research seems to be saying is: "Practice Makes Perfect". This is a deceptively simple answer. During their medical training, radiologists have to learn much more than simply how to read mammograms, and there is not enough time in the radiology residency program to make expert mammographers. Rather, what may be needed is a more effective way to train residents during their clinical residency in mammography. We need to supplement apprenticeship mentoring by expert computer systems. Expert computer systems can provide systematic feedback tailored specifically to each resident's level of training and experience. We propose to use CAVS, which can be "tuned" to provide systematic feedback about regions of the mammogram deemed "suspicious" based on analysis of eye-position dwell times. Prolonged visual dwells will be used to localize image regions for re-evaluation and decision making. Thus, CAVS may hold the key to more effective mammography training.

(10) REFERENCES

1. Krupinski EA, Nodine CF, Kundel HL. Enhancing recognition of lesions in radiographic images using perceptual feedback. Opt Eng 1998;37: 813-818.

2. Nodine CF, Kundel HL, Mello-Thoms et al., How experience and training influence mammography expertise. Acad Radiol. 1999;6:575-585.

3. Kundel Hl, Nodine CF, Krupinski EA. Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. Invest Radiol. 1990; 25: 890-896.

4. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Acad Radiol. 1996; 3: 137-144.

5. Nodine CF, Mello-Thoms, Weinstein SP et al. Blinded review of retrospectively visible but unreported breast cancers: An eye-position analysis. 2000, in preparation.

(11) APPENDICES

1. Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, Conant EF. How experience and training influence mammography expertise. Acad Radiol, 1999; 6: 575-585.

2. Nodine CF, Mello-Thoms C, Weinstein SP et al. Do subtle breast cancers attract visual attention during initial impression? SPIE Medical Imaging 2000; 3981: 156-159.

3. Mello-Thoms C, Nodine CF, Weinstein SP et al. An unobtrusive method for monitoring visual attention during mammogram reading. SPIE Medical Imaging 2000; 3981: 160-163.

4. Mello-Thoms C, Dunn S, Nodine CF et al. Image structure and perceptual errors in mammogram reading: A pilot study. SPIE Medical Imaging 2000; 3981: 170-173.

5. Kundel HL, Weinstein SP, Conant et al. A perceptually tempered display for digital mammograms. Radiographics 1999; 19: 1313-1318.

6. Nodine CF, Mello-Thoms C. The nature of expertise in radiology. In J Beutel, HL Kundel, RL Van Metter (Eds.) Handbook of medical imaging: Vol 1. Physics and psychophysics. Bellingham, WA: SPIE Press 2000; 859-894.

How Experience and Training Influence Mammography Expertise[1]

Calvin F. Nodine, PhD, Harold L. Kundel, MD, Claudia Mello-Thoms, MSEE
Susan P. Weinstein, MD, Susan G. Orel, MD, Daniel C. Sullivan, MD, Emily F. Conant, MD

Rationale and Objectives. The authors evaluated the influence of perceptual and cognitive skills in mammography detection and interpretation by testing three groups representing different levels of mammography expertise in terms of experience, training, and talent with a mammography screening–diagnostic task.

Materials and Methods. One hundred fifty mammograms, composed of unilateral cranial-caudal and mediolateral oblique views, were displayed in pairs on a digital workstation to 19 radiology residents, three experienced mammographers, and nine mammography technologists. One-third of the mammograms showed malignant lesions; two-thirds were malignancy-free. Observers interacted with the display to indicate whether each image contained no malignant lesions or suspicious lesions indicating malignancy. Decision time was measured as the lesions were localized, classified, and rated for decision confidence.

Results. Compared with performance of experts, alternative free response operating characteristic performance for residents was significantly lower and equivalent to that of technologists. Analysis of overall performance showed that, as level of expertise decreased, false-positive results exerted a greater effect on overall decision accuracy over the time course of image perception. This defines the decision speed–accuracy relationship that characterizes mammography expertise.

Conclusion. Differences in resident performance resulted primarily from lack of perceptual-learning experience during mammography training, which limited object recognition skills and made it difficult to determine differences between malignant lesions, benign lesions, and normal image perturbations. A proposed solution is systematic mentor-guided training that links image perception to feedback about the reasons underlying decision making.

Key Words. Breast radiography; education; radiology and radiologists.

One of the outstanding characteristics of an expert in radiology is the speed and accuracy with which he or she decides whether an abnormality is present on a medical image (1–3). Acquiring expertise in radiology requires specialized training, experience, and some degree of talent. How much and what kind of training and experience has been the subject of an organized body of research that has emerged from the field of artificial intelligence (4,5). In this study, we evaluated the influence of perceptual and cognitive skills in mammography detection and interpretation by comparing the performance of experienced radiologists (mammographers), radiology residents, and mammography technologists. The study focused on the performance of the radiology residents, who were receiving training and mentor-guided experiences during mammography rotations that presumably provided a basis for mammography expertise.

It is difficult to find a yardstick to quantify the experience required to achieve expertise in mammography, but one could consider a reading on each case that results in a diagnostic report as a learning-experience trial. This measure of experience ignores immediate feedback, which is important for perceptual learning but is typically absent in clinical practice. In the context of medicine, training consists of mentored experience in which the resident reads medical images and then reviews them with the mentor.

Acad Radiol 1999; 6:575–585

[1] From the Department of Radiology, 308 Stemmler Hall, 36th & Hamilton Walk, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6086 (C.F.N., H.L.K., C.M.T., S.P.W., S.G.O., E.F.C.), and the National Cancer Institute, Rockville, Md (D.C.S.). Received March 19, 1999; revision requested May 5; revision received May 21; accepted May 21. C.F.N. supported in part by USAMRMC grant DAMD17-97-1-7103. **Address reprint requests** to C.F.N.

This training is designed to build feedback into the mentor-guided reading experience, but feedback is neither immediate nor systematic once the resident enters practice. If, for the moment, each read-and-reported case is considered an experience trial, regardless of whether it has been accompanied by feedback, expertise in mammography translates roughly into an average case reading experience equivalent of about 10,000 cases over a period of 3 years (6). This amount of experience compares favorably with estimates of the number of games a chess player plays to reach grand master status (7). The average radiology resident's case reading experience in a mammography rotation over 4 years is about 650 cases, of which only a dozen or fewer may be actual cancers. This means that extensive reading experience after residency is required to reach proficiency as a mammographer. Thus, the amount of experience that a radiology resident receives in training is literally a drop in the bucket.

Visual search is important for detecting lesions in mammograms, but in experts this search skill seems to be specifically tuned for detecting breast lesions embedded in mammograms and does not transfer to similar search tasks in which hidden words and figures are embedded in pictorial scenes (8). It may not even effectively transfer to reading radiographs of areas outside of the breast. Efficient search skills make expert mammographers fast, accurate recognizers, classifiers, and decision makers. Eye-position studies have shown that experts are faster at detecting lesions in chest or breast x-ray images than are less expert observers and that visual-gaze duration (or dwell), which is assumed to reflect visual information processing, is related to decision outcome (6,9). In general, observers dwell longest on the areas in which they report abnormalities, whether their results are true-positive or false-positive. Areas considered negative receive the shortest dwell times. False-negative decisions are the exception. In many instances, observers dwell almost as long on areas containing abnormalities that they report as negative as they do on similar areas that they report as positive, suggesting that they consider these areas to be troublesome even though they reported them as negative.

The fact that cumulative dwell time predicts misses is important in the context of the present study, because it reflects the recognition and decision making that lead up to a diagnostic outcome in much the same way that decision time reflects the gathering of information that leads up to a localization decision. However, visuospatial localization of regions of interest obtained with eye-position recording cannot be derived from decision-time data. The analysis of visual dwell and its relation to information processing leading to a decision outcome suggests that chronometric analysis of the relationship between decision times and decision outcomes may compliment visual dwell data. Experimental psychology has studied reaction time, which is closely related to decision time in the present study, because it "can help one trace the time course of information processing in the human nervous system" (10).

If the goal of mentor-guided experience during resident training is to provide the basis for expertise in mammography, then an important question is: What kind of skills are acquired? Previous research has helped to identify three general areas in which experts skills operate: (*a*) visual search, (*b*) pattern and object recognition, and (*c*) decision making. Because a key characteristic of mammography expertise is the speed-accuracy relationship in decision outcome, the present study focused on how decision making changes as a function of training and experience by comparing groups of observers with different dimensions of speed and accuracy. This comparison entails measuring decision times of observers during mammographic interpretation on a digital workstation and analyzing their decisions by comparing them against a truth table.

Three questions were explored. First, how does performance change as a function of mentor-guided reading experience? Second, how does decision outcome relate to decision time for each decision event during image perception? Finally, what is the likelihood of true versus false decision outcomes over the time course of image perception and decision making? This last question was initially addressed by Christensen et al (11), who were interested in the relationship between what they called "search time" and "perception" in the interpretation of subtle abnormalities and nonpulmonary lesions in chest radiographs. Search time was defined as how long it took to identify an abnormality. Given the possibility of multiple abnormalities per image, there could be multiple decisions per image. Each decision was timed and counted as a decision event. Maximum search time per image was 4 minutes, but most decisions took 1.84–2.68 minutes on average. To compensate for the efficiency associated with faster search times, the actual search time was adjusted by covarying it with the number of decision events within the maximum allotted search time per image. So experienced readers (faculty radiologists) made statistically significantly more decisions in less time than inexperienced readers (radiology residents). By mapping the search times of decision events against a truth table, they were able to plot the time course of true- and false-positive decision outcomes. The analysis of time-

perception data revealed that true-positive results outpaced false-positive results throughout the time course of viewing for experienced readers, whereas false-positive results overtook true-positive results during the time course of viewing for inexperienced readers.

MATERIALS AND METHODS

The mammography test set consisted of craniocaudal (CC) and mediolateral oblique (MLO) paired views from 78 unilateral mammogram cases, for a total of 156 images. The images were digitized (Lumiscan model 100 digitizer; Lumysis, Sunnyvale, Calif) by using a 100-μm spot size. The mammograms were of a single breast and were selected by two mammographers (S.G.O., D.C.S.) from a database of mammography cases taken from the archive of the Hospital of the University of Pennsylvania. These mammographers were later used in the study, but over 2 years had elapsed prior to their testing, and each mammographer contributed only about half of the mammograms to the test set. The mammograms were assembled from cases classified by mammography assessment as normal for at least 2 years, cases classified by mammography assessment as benign and proved by biopsy results to be benign, and cases classified by mammography assessment as malignant and proved by biopsy results to be malignant. The test set contained 25 cases with 15 instances of malignant masses and 14 instances of malignant calcifications shown on both views, one instance of an architectural distortion underlying a malignancy on both views of one breast, and one instance of a single malignant calcification present on only one view. It also contained 24 cases with 12 instances of benign masses and 12 instances of benign calcifications shown on both views and 26 cases considered to be normal. In addition, three practice cases were included: two showing lesions on both views and one lesion-free normal case. For all cases, the two mammographers (S.G.O., D.C.S.) selected mammograms containing subtle benign and malignant lesions. Many of the normal mammograms contained ambiguous image perturbations and thus were considered "difficult normals" by the two mammographers.

The test set was displayed on a single 19-inch, gray-scale monitor (GMA 201, Tektronix, Beaverton, Ore) interfaced to a Sun Sparc 10 computer (Sun Microsystems, Sunnyvale, Calif). At the time of testing, the brightness of the monitor was 127 cd/m^2. This brightness level is low for current state-of-the-art mammography displays and may have led to higher than normal error rates, at least for the inexperienced viewers. Each display consisted of two views of a single breast displayed in the center of the monitor at 2,048 × 2,048-pixel resolution. The gray scale was adjusted for each image by the experimenters (C.F.N., H.L.K.) to a setting that covered the gray-scale range of the breast-only portion of the image. The CC view was shown on the left half of the display screen and the MLO on the right half of the display screen. This is not a typical format for reading mammograms, but we were interested in determining how well observers with different levels of expertise could locate lesions in two views.

Three groups of observers representing different levels of mammography training and reading experience participated: staff mammographers with more than 5 years' experience as dedicated breast imagers (n = 3); 2nd-, 3rd-, and 4th-year radiology residents undergoing a mammography rotation (n = 19); and radiology technologists with 1–9 years' experience in mammographic imaging, but no reading experience (n = 9).

The procedure for testing observers was similar to the interruption technique used by Berbaum et al (12) to obtain response times during visual search. However, the observers viewed the images on a workstation. Lesion identification and decision confidence was entered by "clicking" with a mouse-driven pointer on a menu called up at the time that a lesion was localized. The time from the onset of the display until a decision was made, referred to as decision time, was automatically recorded. The observers were told that they were being tested on their ability to screen for malignancy in a two-view mammographic display of a single breast. If a malignancy was detected, they were to move the cursor to the lesion location and click on it. This action recorded the lesion location and called up a special menu from which they could classify the lesion as a mass, calcification, or architectural distortion and could rate their level of suspicion of malignancy as definitely malignant, highly suspicious for malignancy, moderately suspicious for malignancy, or low suspicion of malignancy. If the observer decided that the two views displayed were free of malignancy, he or she clicked "Return to Routine Screening" on the general menu. If the observer detected a benign lesion, he or she was instructed to treat the mammogram as lesion-free and click "Return to Routine Screening." In addition to these instructions, observers were told to localize malignant lesions on both views, if possible, and to point to the center of masses or a group of calcifications. After three practice trials with the experimenter, to familiarize themselves with the

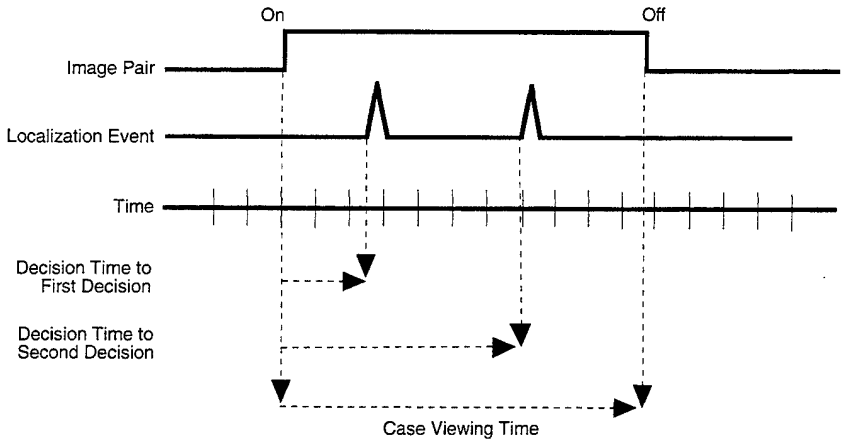


Figure 1. Diagram shows the relationship between image-display presentation and decision events signalled by the observer's clicking the location of a breast lesion on an image with the mouse. Decision time was measured from the onset of the image display to the onset of a decision event. Performance was measured for the task of reading a pair of breast images consisting of CC and mediolateral oblique MLO views. Therefore, more than one decision event was typically timed during each image-display presentation. Offset of the display occurred when the observer clicked on "Next Image."

workstation cursor operations, observers were left to view the 75-case test set on their own. Viewing time per case was unlimited. Decision times were recorded each time a lesion was localized by cursor control, but the observers were not told that their responses were being timed. Because multiple responses were made per two-view image pair, each localization event signaled the occurrence and time of a decision, indicating the presence of a true or false malignant lesion. Figure 1 shows how these events were translated into decision-time measures. For our analysis of decision times, we used the method of survival analysis to generate the cumulative time course of decision outcomes during the time course of viewing. Survival analysis has the advantage of adjusting individual decision times for decision outcomes per case by the total decision-making time required for a case. Thus, our analysis of decision times focused on the cumulative number of decision events per group over the time course of viewing. This is similar to the Christensen et al (11) analysis, which focused on the cumulative number of decision events per group over the time course of viewing 100 chest radiographs.

Analysis of cursor events for localizing, classifying, and rating lesions was accomplished by comparing the observers' decisions against a truth table. The truth table was generated from a combination of mammographic assessment by two of the authors (S.G.O., D.C.S.) and biopsy information on each case. Because all pairs of positive images contained at least two lesions, alternative free response operating characteristic (AFROC) analysis was carried out, treating the pair of positive images as the unit of analysis. This was consistent with the instructions for the task and provided evidence on how well observers with different levels of mammography expertise coordinated lesion localization in a second view, given lesion detection in the first view.

For the AFROC analysis, 30 pairs of malignant lesions were identified as appearing on 25 image pairs. These 60 lesions were counted in the malignant-positive category. The 24 image pairs containing benign lesions plus the lesion-free images (total of 50 image pairs) made up the nonmalignant category. In the AFROC analysis, we counted all correctly localized lesions within ±0.41 cm of the true location on the malignant two-view image pairs (2 standard deviations of mean accuracy of 0.28 cm for mammographers) and counted only the highest-rated false-positive results for the 50 nonmalignant image pairs (equivalent to counting false-positive images; see [13]). It should be noted that this performance criterion ignores classification information that we thought unreasonably stretched the assumptions underlying the two-alternative force choice experimental framework. Basically, the AFROC was designed to measure detection performance. However, because of the importance of the classification decision in mammography, we will provide a separate analysis of the classification data to show how this performance criterion is influenced by the level of expertise.

RESULTS

Overall Performance

Overall detection and localization of breast lesions was assessed as a function of level of expertise. We compared A1, the area under the AFROC curve, for mammographers, residents, and radiology technologists. The AFROC plots the fraction of actual target locations reported (true-positive decisions) against the fraction of images with any false-positive decisions. In our case, we plotted only the highest-rated false-positive decisions on normal or benign images. Figure 2 shows AFROC curves

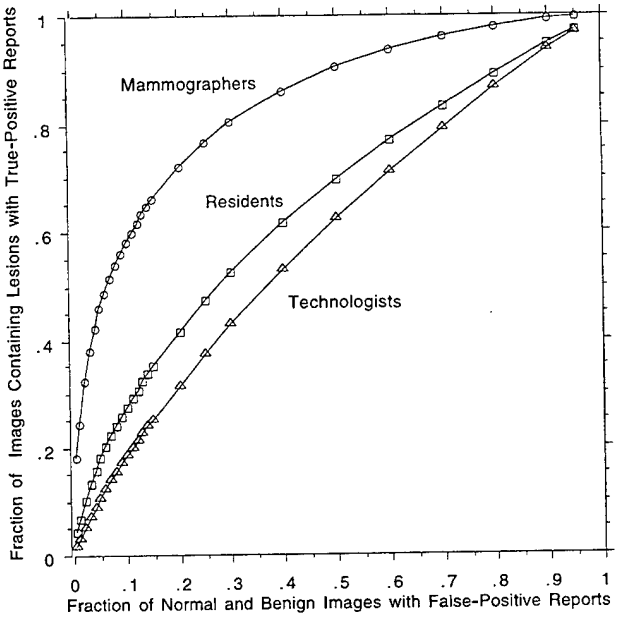


Figure 2. AFROC curves show mean decision accuracy for experienced mammographers (n = 3), radiology residents (n = 19), and mammographic technologists (n = 9). For this analysis, it was assumed that there were 60 malignant lesions on 25 image pairs (CC and MLO views) and 50 malignancy-free images. False-positive results were counted only for malignancy-free images. A computer program called ROCFIT was used to produce an ROC curve after averaging over the confidence intervals for each group of observers. (ROCFIT is part of a set of curve-fitting and estimation programs called ROCKIT, which is available at *http://www-radiology.uchicago.edu/sections* by clicking on "Kurt Rossman Laboratory" and then on "ROC Analysis.")

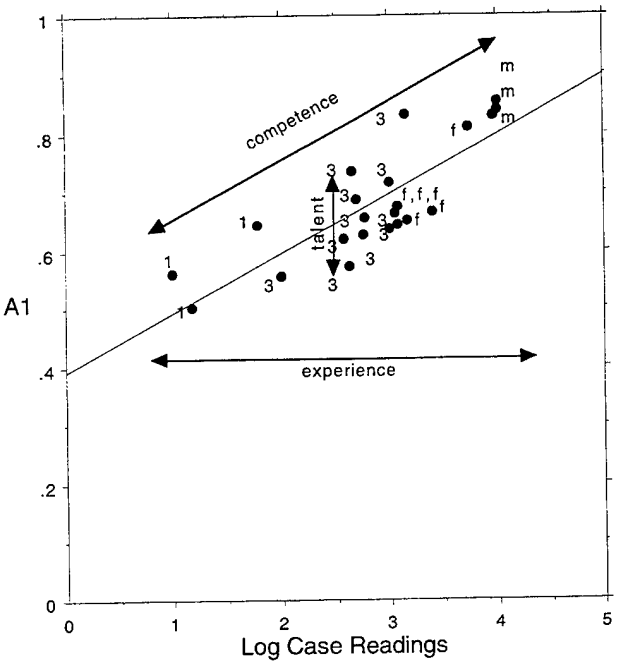


Figure 3. A regression analysis of overall performance measured as A1 as a function of $\log^{10}$ number of cases read over a 3-year period by three experienced mammographers and 19 radiology residents undergoing clinical mammography rotation. When case readings are zero, the regression line intercepts the y axis at A1 = 0.393, which is close to chance performance. With mentor-guided case reading training and experience, A1 performance increases. The numbers and letters within the figure indicate the level of training of the observers: *1* = 1st- and 2nd-year residents, *3* = 3rd- and 4th-year residents, *f* = fellows, and *m* = mammographers.

for the three groups. The average area per observer derived from analysis of variance of A1 values was 0.840 (standard deviation, 0.039) for mammographers, 0.653 (0.058) for residents, and 0.592 (0.062) for technologists. All of these values are above chance performance, which for the AFROC is 0.000. Analysis of variance of A1 values indicated, not surprisingly, that the overall performance accuracy of mammographers was statistically significantly better than that of either residents or technologists, who did not differ from one another ($P < .01$, Scheffe test). By contrasting performance for these groups, which represented different levels of training and experience, we hoped to gain insights into the nature of mammography expertise.

Relation of Case Reading Experience to Development of Mammography Expertise

To provide a clearer picture of how the three groups differ in their experience at reading mammograms, we obtained data on the number of mammographic reports generated by the residents and mammographers. The 19 radiology residents who were part of the study represented mainly 3rd-year (n = 7) and 4th-year (n = 8) residents, plus four fellows who had mammography reading experience varying from 10 to 2,465 cases over a 3-year interval. Over the same period, the three mammographers read 9,459 to 12,145 cases. The relationship between A1 and log number of cases read is shown in Figure 3 for all observers. Figure 3 shows a significant linear-regression fit of the data (R^2 = .667) with a positive slope, suggesting that reading skill, as reflected by A1 performance, increases directly with log case reading experience (F [1,22] = 44.15; $P <$ .0001). The regression line intercepts the y axis at A1 = 0.293, which implies close to chance performance with zero reading experience. A log scale was used to represent the effects of case reading experience because several investigators have suggested the relationship between practice and learning is best expressed by a power function (14). The range of case reading experience in Figure 3 was from 1.0 log case reading to 4.1 log

case readings, or from about 10 to 12,000 cases. Two residents at the beginning of mammography training with little case reading experience performed at an A1 of about 0.500. The fact that their performance is above chance at the beginning of the mammography rotation can be attributed to their talent and their subspecialty training in other areas of radiology. The training levels of the observers are indicated by the numbers or letters associated with the data points. Overall performance increases in an orderly progression with training level.

Identification of Lesions in Two Views

Our hypothesis was that one aspect of performance that might differentiate levels of expertise was how successful observers were at identifying pairs of lesions in a two-view (CC and MLO) display. This hypothesis was based on the assumption that when mammography experts detect a lesion in one view they look for confirmation in a different view. Mammographers talk about using projective geometry principles to predict from a detected lesion to a likely "plane of interest" in which to search for the corresponding "depth" lesion projection. If a detected lesion can be paired in a second view, this provides confirmation that it is a real target. To follow up on this, we analyzed malignant lesions (true-positive decisions) and benign lesions (false-positive decisions) that appeared on CC and MLO views per case by referring to the truth table. The identification of paired localizations on lesion-free areas of images (false-positive decisions) was more speculative, because these were imaginary. To account for paired localizations on lesion-free areas of images (false-positive decisions), we identified sequential decisions— from CC to MLO view or vice versa—that were classified as being malignant and of the same type (eg, mass, calcifications, or architectural distortion). Consistent with the pattern of results in the AFROC analysis, the identification of paired lesions was related to level of expertise. Proportionally more paired lesions were reported and correctly classified by the mammographers than by the residents or technologists. The proportion of correctly paired-lesions was 0.82, 0.56, and 0.50 for mammographers, residents, and technologists, respectively. Proportionally fewer lesions were seen and reported correctly in only one view by all groups, and the corresponding proportions were much lower—0.14, 0.14, and 0.12, for mammographers, residents, and technologists, respectively.

Decision Time and Decision Outcome

The regression plot in Figure 3 shows the relationship between performance and case reading experience. We hypothesized that the decision speed–accuracy relationship, which is a hallmark of expertise, should accompany this improvement in performance, so we looked at decision times as a function of decision outcome, again taking into account that observers were interpreting an image pair containing CC and MLO views and thus possibly making two or more decisions per case. To identify the sequencing of decisions per case, the paired decisions were broken down into those occurring in the CC view on the left side of the display and the MLO view on the right side of the display. For these paired decisions, decision times to the first decision were inversely related to level of expertise, with mammographers significantly faster than residents ($P < .01$, Scheffe test) and residents significantly faster than the technologists ($P < .0001$, Scheffe test). For mammographers compared with residents, 32% more of their initial responses were true-positive, and these initial responses were reported faster than those of residents. Mean decision time for the first correct decision per pair was 15.66 seconds versus 21.56 seconds (t [376] = 3.91; $P < .001$). Technologists detected fewer true-positive results and took even longer to decide (28.08 seconds). Decision time was also inversely related to level of expertise in a similar pattern for classification of localized lesions. Mammographers correctly classified 38% more lesions and did so faster than residents ($P < .05$) and technologists ($P < .001$). Mean decision time for mammographers was 16.51 seconds for classifying masses and 19.77 seconds for classifying calcifications. Both of these findings support the decision speed–accuracy relationship associated with expertise.

Finally, to provide some perspective on how true-positive results related to false-negative results, we looked at decision times when all lesions were completely missed on images containing malignant lesions. In this case, total image duration was assigned as the decision time. This result might be considered a "clean" miss in that no lesion was reported, even though a lesion was present during the entire time that the image was examined. Of 579 total false-negative decisions, 51% were clean misses. Mean decision times differed little for the clean-miss false-negative category, as they ranged from 38 to 46 seconds. However, standard deviations of the mean decision times ranged from 4.6 seconds for mammographers to 41.6 and 52.5 seconds for residents and technologists, respectively. These values indicated that the latter two groups had considerable indecision about not making a positive report after examining two views of an image containing a truly malignant lesion. The range of mean decision times for clean misses

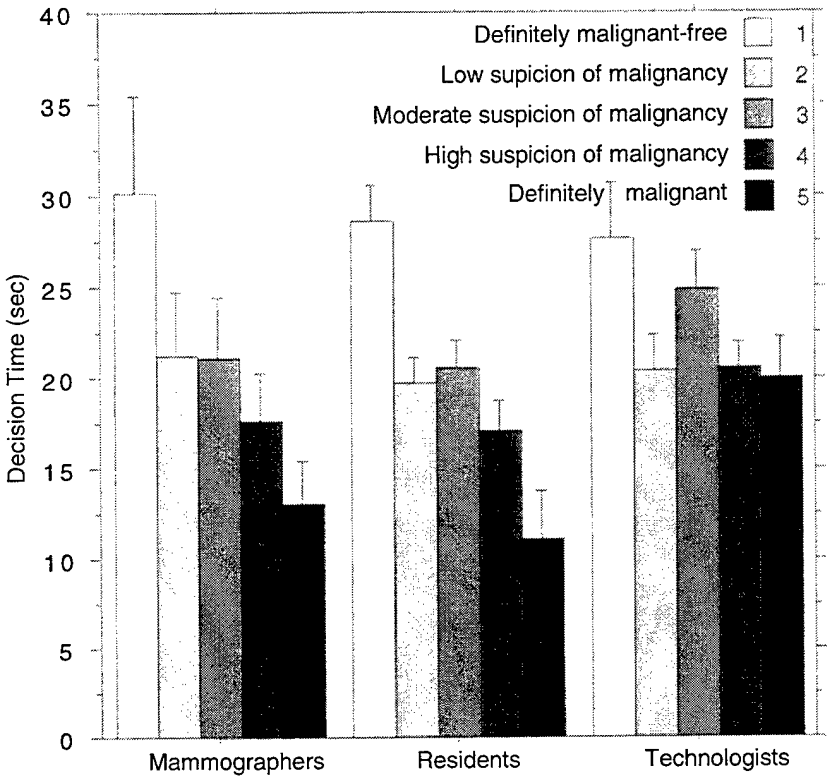


Figure 4. Decision time as a function of decision-confidence ratings for mammographers, residents, and technologists. A confidence rating of 5 indicated the lesion was definitely malignant; 4, highly suspicious for malignancy; 3, moderately suspicious for malignancy; 2, low suspicion of malignancy; and 1, definitely malignant-free.

was longer than that of any other decision outcome categories and seems to complement the finding obtained from monitoring eye position of prolonged visual dwell for false-negative decisions. Observers spent a longer time deciding to call a positive case negative. Overall, clean-miss false-negative decisions took significantly longer than true-negative decisions (t [864] = 4.22; P <.001). Of course, we cannot confirm that the true lesions were actually scrutinized from the decision time data, but the long decision times and wide variances suggest much uncertainty surrounding decision making.

Relationship of Decision Time to Use of Confidence Ratings

The similarity of the relationship of decision outcome to decision time for mammographers and residents suggests that they may be using similar underlying detection and decision strategies. One measure that reflects underlying decision strategy is how observers used the confidence ratings in making decisions. It is reasonable to assume that the more sure observers are that they have detected a lesion, the faster they are at making a decision. Figure 4 shows the relationships between decision time and use of confidence ratings for the three levels of expertise. The general pattern for the mammographers and residents was that decision times were inversely related to the confidence rating. The longest decision times were for definitely lesion-free images (rating = 1), and the shortest decision times were for definitely malignant image locations (rating = 5). This pattern suggests that both groups had a similar perceptual thresholding basis for the decision, which is an important factor in developing a decision-making strategy. The pattern for technologists showed virtually no relationship between decision time and use of confidence ratings. Only confidence 1 ratings were prolonged. No evidence showed that decision times were faster when technologists were more confident that a malignant lesion was present on an image.

Time Course of Decision Outcomes

So far, two interesting generalizations come out of the decision time analysis. First, the decision speed–accuracy relationship was found to be related to level of expertise. Figure 5 summarizes the decision speed-accuracy relationship expressed by d' (cumulative) as a function of viewing

time for mammographers, residents, and technologists. Cumulative values for true-positive and false-positive decisions to both normal and benign images on a per case basis (paired decisions) as a function of decision time were obtained from survival analysis. These values were then transformed using the formula $d' = z$ (true-positive decisions/30) $- z$ (false-positive decisions)/50, where z can be interpreted as a deviate of the unit normal curve. This formula can be thought of as correcting the true-positive fraction by the false-positive fraction. Decision accuracy consists of detecting perturbations in images, testing them for signs of malignancy, and correctly classifying them as masses, architectural distortions, or calcifications. This complex decision requires discriminating malignant from benign lesions, and malignant from normal anatomic variants in the breast image. Decision accuracy can be expressed as A1, the area under the AFROC curve, or as d', the index of detectability derived from the true-positive fraction and the false-positive fraction at a specific decision threshold, as shown in Figure 5. Looking at performance in this way shows clear differences as a function of level of expertise.

Second, decision times were longer for false than for true decision outcomes. To consider whether these false decisions tended to occur early or late in the time course of image perception, we looked at both paired and single decisions. A paired decision is one in which the observer sequentially localized a suspected lesion (true or false) on both CC and MLO views. Figure 6 shows the mean number of paired true-positive decisions and paired false-positive decisions for normal regions of the images and benign lesions for mammographers, residents, and technologists as a function of viewing time per case. Figure 7 shows the same plot for single decisions, as contrasted with paired decisions. The most striking feature of Figure 6 is the technologists' high rate of false-positive results for normal regions in relation to the rate of their true-positive results, for paired decisions. In Figure 7, it is the high rate of false-positive results for normal regions for all groups for single decisions.

These plots show that for mammographers the rate of true-positive decisions for normal regions is faster than the rate for false-positive decisions, but false-positive decisions for normal regions continue to plague performance throughout the time course of viewing. False-positive decisions for benign lesions drop out relatively early; thus, overall performance continuously improves with decision time until about 60 seconds. Perhaps our mammographers should have considered stopping at this point, because

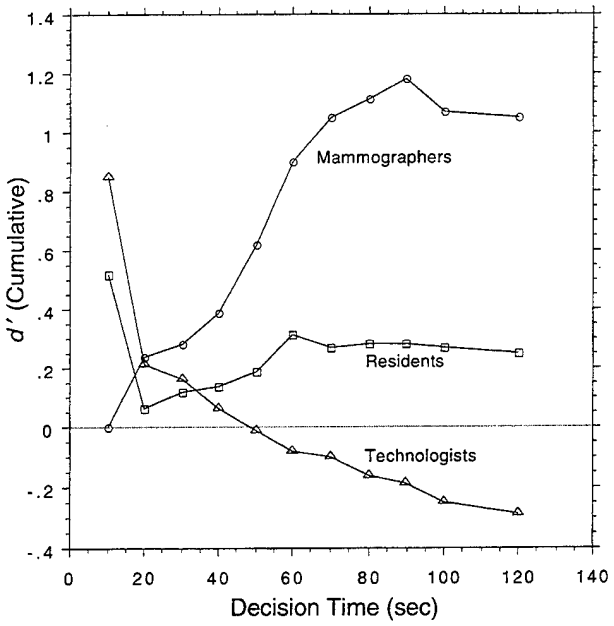


Figure 5. Speed-accuracy relationship indicated by d' as a function of decision time for mammographers, residents, and technologists. Overall performance measured by d', which is the normal deviate (z) of true-positive results minus the false-positive results, increased for mammographers and to a lesser extent for residents. Overall performance decreased below chance ($d' = 0$) for technologists, which means that false-positive results outnumbered true-positive results.

false-positive decisions for normal regions increased faster than true-positive decisions. The rate of true-positive decisions is slower for residents because of continuous competition from false-positive decisions for normal regions up to 60 seconds. As with mammographers, the false-positive decisions for benign regions peak earlier. The technologists show a decrease in overall performance over time because they continued to make more false-positive decisions for normal regions than true-positive decisions.

DISCUSSION

Understanding the Nature of Expertise

The goal of the present study was to understand better the nature of expertise in mammography. Expertise in mammography, as we have defined it here, refers to diagnostic performance skills that enable the observer to localize a breast lesion and correctly decide that it is or is not malignant on the basis of two views. Admittedly, our task was somewhat artificial in the sense that we mixed lesion detection, which is the focus of mammography screening, with diagnostic interpretation, which is the focus of diagnostic follow-up. The next step is to break the task apart and do a

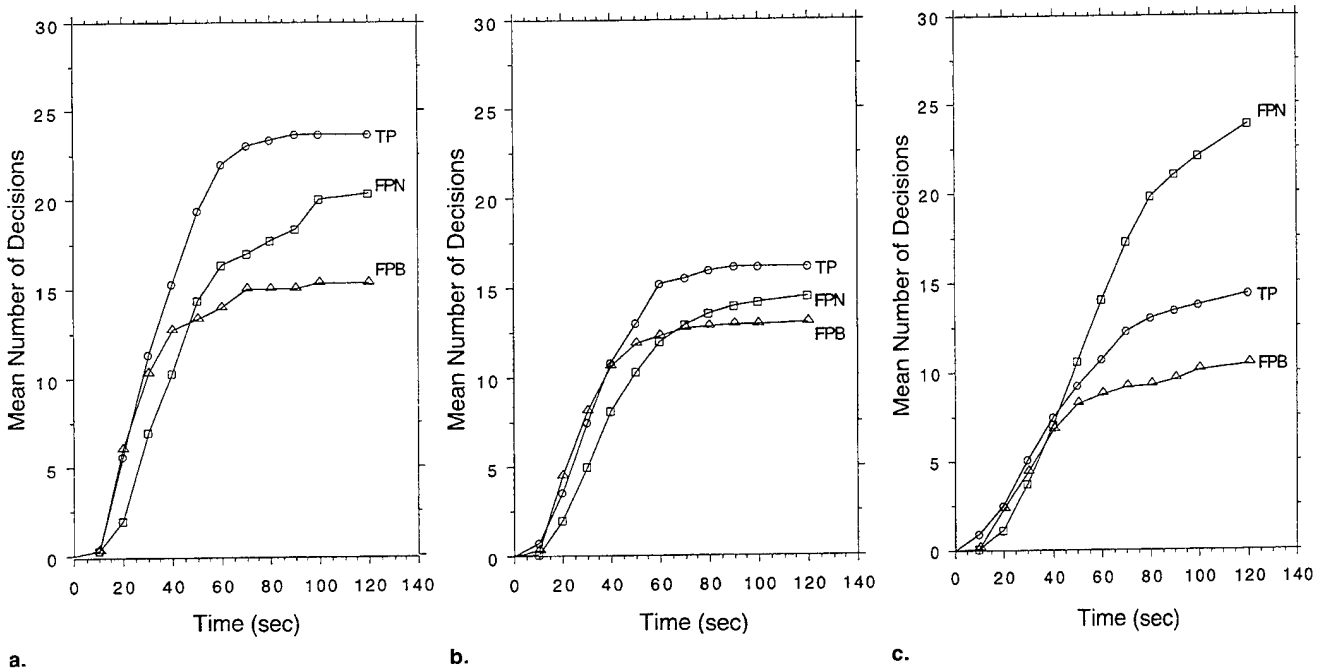


Figure 6. Cumulative mean numbers of paired decisions per case as a function of the decision time course of viewing for true-positive (*TP*) decision outcomes, false-positive decision outcomes on normal images (*FPN*), and false-positive decision outcomes on images containing benign lesions (*FPB*) for **(a)** mammographers, **(b)** residents, and **(c)** technologists. Paired decisions were measured. Of the malign cases, all but one contained lesions in both CC and MLO views. As this figure indicates, within 60 seconds, the mammographers had localized 23 (92%) of 25 paired true lesions.

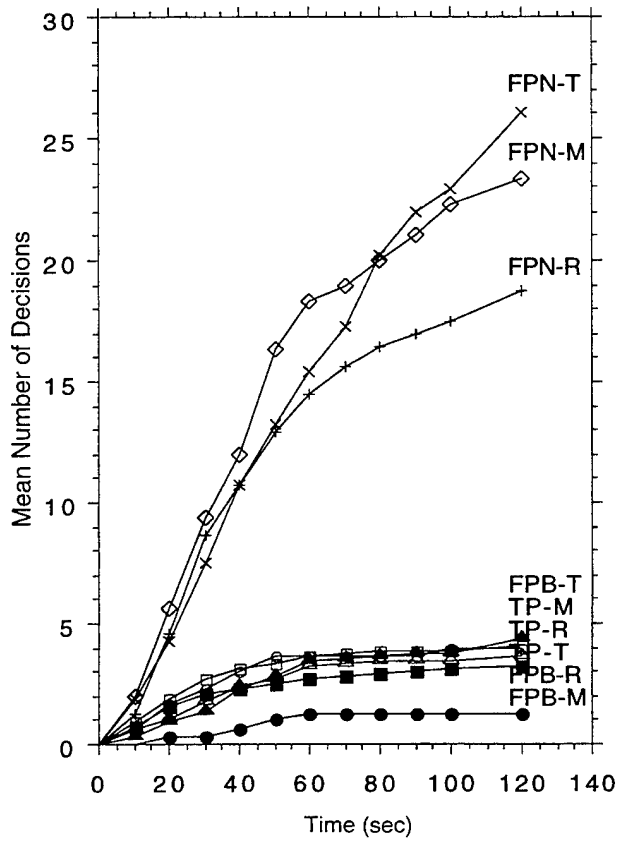


Figure 7. Cumulative mean number of single decisions as a function of the decision time course of viewing for true-positive (*TP*) decision outcomes, false-positive decision outcomes on normal images (*FPN*), and false-positive decision outcomes on images containing benign lesions (*FPB*) for mammographers (*M*), residents (*R*), and technologists (*T*).

two-part test, which will come closer to the American College of Radiology Breast Imaging Reporting and Data System format. Moreover, even though the diagnostic skills that we have studied are an essential part of mammography diagnosis, the study is limited, as only CC and MLO views were shown with no capability for prior studies, additional views, or magnification views. Additional special mammographic images, such as spot compression or magnification views, and breast ultrasound imaging, both of which are important parts of mammography expertise, were not employed in the present study. On the basis of the information these modalities provide, the mammographer may decide that the finding is normal, benign, or probably benign but recommend short-term follow-up or a biopsy.

Why Are Experts Faster and More Accurate?

Our analysis has related A1 and d', measures of overall performance, to decision time to shed light on basic

perceptual and decision-making skills. Differences in speed and accuracy between mammographers and residents seem to be related to the experience required to gain expertise, as shown in Figure 3. This suggests that experts are perceptually more sensitive in recognizing lesions than are those with less expertise because the experts have read more mammogram cases, seen more lesions, and differentiated more lesions into malignant and benign categories. In practical terms, this means that through massive amounts of experience experts became perceptually tuned to recognizing familiar common breast structures and detecting odd or novel variations in them. Three to 5 years of dedicated experience reading mammograms affects perceptual learning by exposing mammographers to a wide set of breast image configurations that represent most variations of normality and abnormality. We hypothesize that this concentrated case reading experience with mammographic images has an effect on perceptual learning by producing enhanced recognition skills akin to those attributed to chess grand masters who, according to one estimate, are capable of recognizing on the order of 50,000 different chess configurations (7). It is unclear whether enhanced object-recognition skill is the result of the development of what the artificial intelligence community refers to as "chunking" or template-retrieval structures that aid short-term and long-term memory (14) or, as we have suggested, more critically tuned visual recognition as the result of learning and refining distinctive-feature information used to recognize deviations from prototypic normal breast structures (15,16).

Supporting the argument in favor of the tuning of visual recognition, Sowden et al (16) have shown that massed practice detecting calcifications in positive-contrast mammograms (bright targets on a dark background) positively transfers to a new task in which the calcifications are displayed in negative-contrast mammograms (dark targets on a bright background). This suggests that perceptual learning improves perceptual sensitivity in the detection of low-contrast targets. Massed practice was defined as a detection trial followed immediately by feedback about the correctness of an observer's response. This improvement in perceptual sensitivity occurred even though the amount of massed practice was limited to 720 trials, followed by the transfer test. The key to improvement may be the feedback. Generalizing the results of Sowden et al (16), one cannot help but wonder if the effects of reading experience would be facilitated by computer-assisted visual feedback about decision outcomes delivered for some subset of test cases in which truth could be verified or, at least, agreement consensus reached. The purpose of systematic visual feedback is to make image perception and decision making an integral part of a perceptual-learning reading experience (6,17).

Expertise: Chest Radiology Compared with Breast Radiology

In interpreting performance differences, we have to be careful to separate studies of expertise in chest radiology from those in mammography, because chest radiology studies have emphasized the importance of input from peripheral vision in detecting pulmonary lesions. Peripheral vision is important during the search for inconspicuous pulmonary lesions because a chest radiograph contains so many anatomic landmarks (eg, heart, ribs, lungs, diaphragm). It has been suggested that these anatomic landmarks act as a map, helping peripheral guidance of search (18). Anatomic landmarks are few in the breast (eg, nipple and pectoralis muscle), and breast structures that might serve as landmarks (eg, blood vessels and ducts) are interwoven into the breast image, creating texture differences that are probably too subtle to be selected by peripheral vision during a search. As a consequence, rather than landmarks, we believe that perturbations in parenchymal structure caused by compression of the breast during imaging and desmoplastic reaction from a growing tumor provide focal points of interest during a visual search. The superimposition of parenchymal structures tends to make them visually conspicuous. Because the superimposition of parenchymal structures has the potential to mimic breast lesions, they may be detected by peripheral vision during the initial global survey, scrutinized during subsequent focal scanning, and falsely reported as true lesions. In the detection of breast lesions, it is not only important for the observer to recognize familiar features in the image but also to recognize odd or novel features, examine these in detail (as reflected by fixations and decision time), and weigh their importance in making a decision (6,19). We assume that dwell time spent fixating the lesion, like time spent examining the image prior to making a decision, represents the information processing time required to make a decision.

Decision Strategies

Our study has shown that residents develop decision-making strategies that are similar to those of experts. From a practical standpoint, this suggests that resident training in mammography is effective in providing a general framework for learning radiology image-perception skills. However, residents are not as good as experts at

identifying true breast lesions. We hypothesize that this weakness is due primarily to the lack of fine-tuned visual recognition skills. Because feedback is recognized as a critical part of the reading experience, built into the clinical mammography rotation, it is tempting to speculate that providing computer-assisted feedback training might facilitate visual recognition skills and bring resident overall performance closer to that of their mentors. Despite their limited perceptual experience, many of the radiology residents will join clinical practices and read mammograms as practicing radiologists. Does this mean that the diagnostic performance of practicing radiologists will suffer as a result? Probably, because the overall average performance of residents in the present study had an average receiver operating characteristic curve area of 0.743, which was 12% lower than the national average of 0.845 for 108 U.S. radiologists, assuming approximately the same level of case difficulty for the two test sets (20).

Finally, we have shown that decision accuracy is directly related to amount of case reading experience. At the present time, many radiology departments keep track of the number of cases read by radiologists and residents, yet no recommendations have been proposed as standards.

Our data support the need for minimum requirements in number of case readings, such as those proposed by the latest Food and Drug Administration regulations. As of April 28, 1999, this requirement was 240 case readings within the past 2 years of residency. In addition, we believe that some less abrupt transition between residency and practice (for example, double-reading experience during the 1st year of practice) would greatly improve performance standards (21).

REFERENCES

1. Lesgold A, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y. Expertise in a complex skill: diagnosing x-ray pictures. In: Chi MTH, Glaser R, Farr MJ, eds. The nature of expertise. Hillsdale, NJ: Lawrence Erlbaum, 1988; 311–342.
2. Kundel HL, La Follette P. Visual search patterns and experience with radiological images. Radiology 1972; 103:523–528.
3. Parasuraman R. Effects of practice on detection of abnormalities in chest x-rays. Proc Hum Factors Soc 1986; 309–311.
4. Newell A, Simon HA. Human problem solving. Englewood Cliffs, NJ: Prentice-Hall, 1972.
5. Chi MTH, Glaser R, Farr MJ. The nature of expertise. Hillsdale, NJ: Lawrence Erlbaum, 1988.
6. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Acad Radiol 1996: 3: 1000–1006.
7. Chase WG, Simon HA. Perception in chess. Cognitive Psychol 1973; 4:55–81.
8. Nodine CF, Krupinski EA. Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. Acad Radiol 1998; 5: 603–612.
9. Kundel HL, Nodine CF, Krupinski EA. Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. Invest Radiol 1990; 25:890–896.
10. Posner MI. Chronometric explorations of mind. New York, NY: Oxford, 1986; 218.
11. Christensen EE, Murry RC, Holland K, Reynolds J, Landay MJ, Moore JG. The effect of search time on perception. Radiology 1981; 138:361–365.
12. Berbaum KS, Franken EA, Dorfman DD, et al. Time course of satisfaction of search. Invest Radiol 1991; 26:640–648.
13. Chakraborty DP, Winter LHL. Free-response methodology: alternative analysis and a new observer-performance experiment. Radiology 1990; 174:873–881.
14. Gobet F, Simon HA. Templates in chess memory: a mechanism for recalling several boards. Cognitive Psychol 1996; 31:1–40.
15. Myles-Worsley M, Johnston WA, Simons MA. The influence of expertise on x-ray image processing. J Exp Psychol Learn Mem Cogn 1988; 14:553–557.
16. Sowden P, Davies I, Roling P. Perceptual learning of the detection of features in x-ray images: a functional role for improvements in adults' visual sensitivity? J Exp Psychol Hum Percept Perform (in press).
17. Anderson JR. Cognitive psychology and its implications. 4th ed. New York, NY: Freeman, 1995.
18. Kundel HL, Nodine CF, Thickman D, Toto L. Searching for lung nodules: a comparison of human performance with random and systematic scanning models. Invest Radiol 1987; 22:417–422.
19. Ullman S. High-level vision: object recognition and visual cognition. Cambridge, Mass: MIT Press, 1996; 161.
20. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Med 1996; 156:209–213.
21. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996; 3:891–897.

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Vol. 1, No. 26
ISSN 1605-7422

Reprinted from

Medical Imaging 2000

Image Perception and Performance

16–17 February 2000
San Diego, California

Proceedings of SPIE
Volume 3981

Do Subtle Breast Cancers Attract Visual Attention During Initial Impression?

Calvin F. Nodine, Claudia Mello-Thoms, Susan P. Weinstein, Harold L. Kundel, and Lawrence C. Toto
University of Pennsylvania, Philadelphia, PA 19104-6086

ABSTRACT

Women who undergo regular mammographic screening afford mammographers a unique opportunity to compare current mammograms with prior exams. This comparison greatly assists mammographers in detecting early breast cancer. A question that commonly arises when a cancer is detected under regular periodic screening conditions is whether the cancer is new, or was it missed on the prior exam? This is a difficult question to answer by retrospective analysis, because knowledge of the status of the current exam biases the interpretation of the prior exam. To eliminate this bias and provide some degree of objectivity in studying this question, we looked at whether experienced mammographers who had no prior knowledge of a set of test cases fixated on potential cancer-containing regions on mammograms from cases penultimate to cancer detection. The results show that experienced mammographers cannot recognize most malignant cancers selected by retrospective analysis.

Keywords: Visual attention, Missed cancers, Retrospective analysis, Eye fixations

1. INTRODUCTION

Should detected breast cancers that can be seen retrospectively on the immediately prior mammogram be considered missed or incident cancers?

This is a difficult question to answer because perceptual knowledge of lesion features and location bias the observer's interpretation in retrospectively looking for the cancer on the prior mammogram. The issue of missed cancers is a major source of malpractice lawsuits filed against radiologists (Berlin, 1999). This is in spite of the fact that even an "expert" making a retrospective analysis cannot neutralize apriori knowledge in viewing the radiographic image after having once recognized the cancer (Berlin, 1996).

Our experiment looked at subtle cancer cases. These consisted of: (a) a group of 20 subtle cancer cases that were not reported on the mammogram immediately prior to detection (mean interval = 14.25 mo.), but were visible in retrospect when analyzed by an experienced breast imager (SPW); (b) a group of 10 true incident cancer cases, and, (c) 10 cancer-free cases (2-year follow up). These subtle cancer cases were digitized to 50 micron image resolution. The image gray scale for each view was automatically set by a linear Look-Up-Table (LUT) algorithm in which a binary version of the original image was used to find the breast outline, and then the intensity range within the original breast image segment was sampled to define the LUT .

In order to design a fair test of the question, we needed to choose observers who were experienced mammographers, but who did not have apriori knowledge of the mammogram cases in the test set. They were, however, given information indicating that they would be seeing subtle lesions, so their suspicion was raised. In addition, we monitored eye position during initial interpretation of the mammograms in order to provide an objective measure of whether or not the subtle lesions were looked at (fixated) independently of being reported. When the initial interpretation was concluded, the observer gave a general impression (normal or abnormal). Without interruption, the observer was given additional viewing time to examine the mammogram case using full-resolution digital zoom. If a potentially malignant or suspicious lesion was recognized, the observer localized it with a mouse cursor. This action called up a menu prompting the observer to classify the lesion by type and give a decision-confidence rating. If no suspicious findings were found, the observer terminated the trial by calling up the next image. This resulted in a default normal decision.

The focus of my paper is on how analysis of eye-position data are related to whether or not subtle lesions are fixated long enough for the observer to make a decision about them, and how these data are related to initial decision outcome and subsequent zooming analysis. The complementary paper to this (3981-25) presented by Claudia Mello-Thoms will focus on how zooming data are related to localizing subtle lesions and how these data are related to eye position and final diagnostic decision outcomes. It should be noted that these two papers are based on the same experiment.

In *Medical Imaging 2000: Image Perception and Performance*, Elizabeth A. Krupinski, Editor, Proceedings of SPIE Vol. 3981 (2000) • 1605-7422/00/$15.00

2. MATERIALS AND METHODS

We recorded eye-position data (ASL, Model 4000SU, Bedford, MA) on 4 experienced breast imagers viewing a test set consisting of 20 retrospectively visible cancer cases not reported on initial screening (NR), 10 prospectively reported cancer cases (R), and 10 cancer-free cases. Two mammographic views, CC and MLO, were digitized for each case and displayed on a 21" high-resolution (2560 x 2048) workstation (Orwin, Model DS5000L, Amityville, NY). This was no ordinary workstation in that a data record was generated on each observer which contained: event times of mouse clicks indicating decision events; lesion locations; eye-fixation locations; zoom locations; and, zoom durations for each mammographic view of each case.

3. RESULTS

Did experienced breast imagers look at subtle lesions long enough to recognize malignancy? We used 1000 ms as the dwell threshold for recognizing a breast lesion based on earlier work in which we showed that a minimum of 1000 ms was required for a positive decision (Krupinski, Nodine, Kundel, 1998).

Considering that NR lesions are true cancers, the answer to the question that prompted this study is "yes". Initially, 66 % of NR lesions v.60 % of R lesions were fixated for >1000 ms.

Phase 1 time was highly correlated with total number of fixation clusters as shown in Figure 1 which relates total number of cumulative clusters per image to phase 1 viewing time. This suggests that most visual search time was spent focally searching and examining image features for possible lesions. This is the effect of "zooming" with the eye.

Insert Fig. 1 here

How does fixating relate to initial decision outcome? Initially, observers over reported as positive 69% of NR and 81% of R test cases. Bar Graph 1 shows the yield of decision outcomes resulting from the initial decision for fixations >1000 ms. for NR and R test cases. Only slightly more than half of NR cases (58%) were correctly interpreted compared 76% of R cases based on initial decision.

Insert Bar Graph 1 here

False positive rates of 28% and 19% are not too far out of line given that in clinical practice, for patients recommended for biopsy, only 1 in 3 will typically have a cancer. But we did not allow observers to perform additional imaging evaluations in the present study.

How long did observers fixate to generate a decision? Observers were suspicious since they were told that they were looking for subtle lesions. Initially, experts eyes fixated subtle lesions, but they had difficulty recognizing true from false malignant lesions. In reality mammographers do not rely on 2 mammographic views alone to determine malignancy, but follow up with additional evaluation images such as mag views, ultrasound and ultimately biopsy.

Mean fixation cluster dwell times by decision outcome for NR and R test cases are shown in Bar Graph 2.

Insert Bar Graph 2 here

Interestingly, decisions with mean dwell times >1000 ms. (n=166) were 7 times longer than the corresponding decisions with mean dwell times <1000 ms. (n= 204).These latter decision times ranged from 372-680 ms. suggesting that 1000 ms. is a good dwell threshold for defining "directed attention".

Bar Graph 2 is based on a lesion analysis of CC and MLO views using truth table generated by the breast imager (SPW). The long dwells, especially for NR test cases, suggest difficulty differentiating true from false malignant lesions, implying a low signal-to-noise ratio. These average dwell times are consistent with previous studies (e.g. Krupinski, Nodine, Kundel, 1998)

Does fixating a potential lesion result in subsequent zooming of it? Yes, 80% of initially fixated lesions were subsequently zoomed. No difference between NR and R.
Fixations that were subsequently zoomed resulted in longer dwells (1884 ms) than fixations that were not subsequently zoomed (1224 ms, $p<.05$, Scheffe test) indicating that findings that captured visual attention were followed up by zooming.

4. DISCUSSION

Experienced breast imagers with high suspicion initially failed to recognize 42% of retrospectively visible subtle malignant breast lesions. Does this mean that these subtle lesions should not be considered "missed cancers" but rather true incident cancers because they could not be differentiated from normal background structures? Probably not.

We have acknowledged the high rate of false positives in this study and attributed it, in part, to increased suspicion on the part of the observers. It is also due to the scoring of overall performance which was done on a lesion basis meaning that observers could, and did, generate FPs on both CC and MLO views. They also got credit for finding cancers on both views. But, from a clinical standpoint, the troubling aspect of this performance was not the high false positive rate, but the higher miss rate.

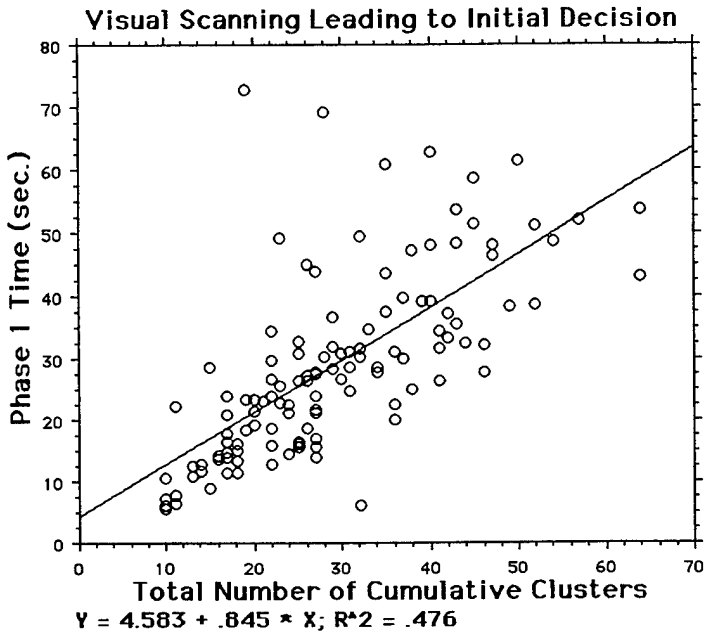
Why most cancers in the NR cases were not recognized at the initial viewing is unclear, but three thoughts come to mind. First, mammographic images are far from perfect and this study used digitized versions which may have degraded the signal-to-noise ratio. However, although subtle, the lesions were retrospectively visible on the digitized version. Second, we did not give the mammographers the option to further evaluate the areas of potential cancer. They knew that they had to rely on the 2 views supplemented only by full-resolution zooming for a malignant/non-malignant interpretation. This is not the way experienced mammographers work in practice and may have played out by a higher than normal miss rate. Finally, I hope this study puts a nail in the "retrospective analysis" coffin. Although a lesion may be visible in retrospect, our experienced breast imagers had extreme difficulty differentiating true positive lesions from false positive ones, even though they were alerted to the presence of "subtle" cancers. So much for expert testimony based on retrospective analysis. It is easy to detect a subtle cancer with the benefit of apriori knowledge, but without it even highly experienced breast imagers stumble.

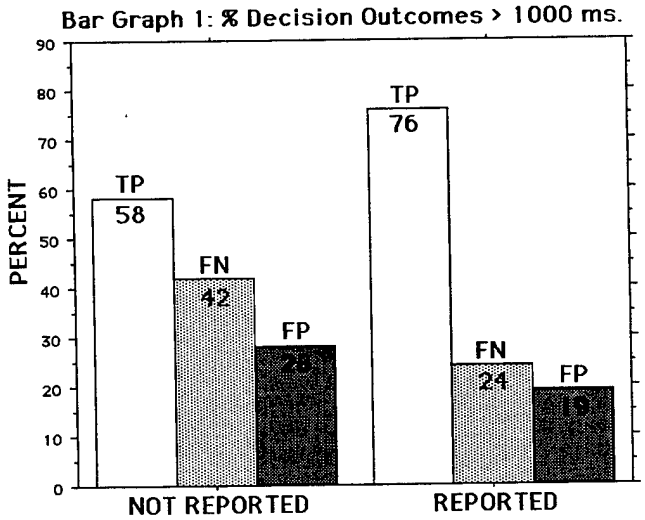
5. ACKNOWLEDGEMENTS

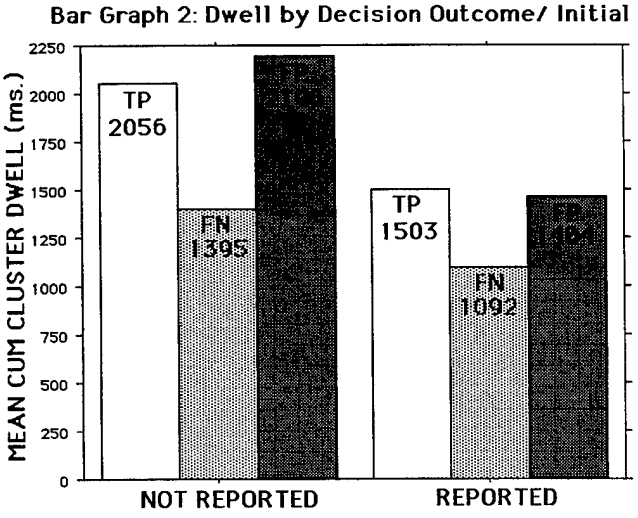
This research was supported in part by DAMD17-97-1-7130

6. REFERENCES

1. Berlin L. The missed breast cancer: Perceptions and realities. ARJ 1999;173:11-61-1167.
2. Berlin L. Perceptual errors. ARJ 1996;167:587-590.
3. Krupinski EA, Nodine CF, Kundel HL. Enhancing recognition of lesions in radiographic images using perceptual feedback. Opt Eng 1998;37: 813-818.


Visual Scanning Leading to Initial Decision
Phase 1 Time (sec.)
Total Number of Cumulative Clusters
Y = 4.583 + .845 * X; R^2 = .476


Bar Graph 1: % Decision Outcomes > 1000 ms.
PERCENT
TP
58
FN
42
FP
TP
76
FN
24
FP
NOT REPORTED
REPORTED


Bar Graph 2: Dwell by Decision Outcome/ Initial
MEAN CUM CLUSTER DWELL (ms.)
TP
2056
FN
1395
TP
1503
FN
1092
NOT REPORTED
REPORTED

3.

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Vol. 1, No. 26
ISSN 1605-7422

Reprinted from

Medical Imaging 2000

Image Perception and Performance

16–17 February 2000
San Diego, California

Proceedings of SPIE
Volume 3981

An Unobtrusive Method for Monitoring Visual Attention During Mammogram Reading

Claudia Mello-Thoms, Calvin F. Nodine, Susan P. Weinstein, Harold L. Kundel and Lawrence C. Toto
University of Pennsylvania School of Medicine, Philadelphia, PA 19104

Abstract

The use of feedback to the observer of the regions of the image that attract prolonged visual dwell (> 1000 ms) has been shown to improve nodule detection performance in reading chest x-rays. The application of such a feed back mechanism in mammography seems appropriate, but it is often discouraged by the inherent difficulties of using an invasive eye-tracking system. In this paper we discuss the use of an alternative method, namely, a digital zoom window, to monitor where the observer's attention is focused on the image. We have shown that the order in which the zooms occur, as well as the duration of certain zooms, is statistically correlated with decision outcome for a given region of the image. Furthermore we show a strong correlation between zooming and prolonged fixation.

Keywords: Breast cancer, eye-position monitoring, zoom window, decision outcome.

1. Introduction

Mammography is the standard screening test for breast cancer. Nonetheless, sensitivity of Mammography is about 85-90%. A question that naturally arises is: these cancers were missed due to faulty search or recognition failure?

Eye position studies have shown that the majority of missed cancers are in fact looked at [1], and the dwell times on these locations are almost as long as on the cancers that are reported. Furthermore, eye position and a dwell time threshold have been used to provide feedback to observers about the locations of possibly missed nodules in chest x-ray readings, and detection performance has improved as a result [2]. Thus, in order to improve breast cancer detection, one interesting alternative is the application of perceptual feedback. Unfortunately, eye position monitoring, using an eye-tracker, is a cumbersome and intrusive research tool. It suffers from a variety of drawbacks, such as the need to keep the calibration updated, adjust for spurious reflections from the observer's eye glasses or contact lenses and reflections from the observer's skin, difficulty in tracking the pupil if the observer tends to lower his or her eyelids, etc. Furthermore, even with a perfect observer there is still some discomfort due to eye dryness and headaches caused by the infrared beam used to monitor limbus reflections. Thus, it is impractical to use such a system for long-term monitoring of the observers' attention when reading medical images.

In their daily practices mammographers read mammograms using a two-pass strategy. In the first pass they globally search the mammogram for typical abnormalities, and in the second, using a magnifying lens, they repeat the search, looking for microcalcifications or other subtle findings. In this way, we hypothesized that by allowing them to use a digital zoom, when reading mammogram cases on a computer workstation, we would be able to monitor where on the image their attention was directed without using any eye position monitoring. Furthermore, we hypothesized that the length of the time that the zoom window is stationary at a particular location is related to the decision outcome yielded at that location, just as the visual dwell is related to decision outcome.

In this paper we will examine the use of this digital zoom window to monitor the experts' visual attention, and compare the results with an eye-tracking system. We will show how the use of the zoom is related to the decision outcomes in a task where the observers are instructed to search for malignancy. We will also show which percentage of these responses had initially attracted the attention of the observers, during a scanning phase in which eye position was monitored using an eye-tracking system, and how many of them were further investigated on a second phase in which the observer was allowed to zoom in onto a region of interest in the image.

2. Materials and Methods

Four experienced observers (2 staff mammographers and 2 fellows undergoing training at the Hospital of the University of Pennsylvania) examined 40 two-view mammogram cases on a digital workstation. These 40 cases were obtained from the archives of the same hospital by one of the authors (SPW), who is a mammographer but did not participate in the study. These cases contained 10 cancer-free cases which had been stable for a period of 2 years (N); 10 cases in which the malignant lesion present had been reported (R), and 20 cases in which a malignant lesion, albeit present and visible, had not been reported, being only found retrospectively (NR). Malignancy for all lesions was determined through biopsy.

The films were digitized using a Lumiscan Model 100 digitizer (Lumisys Inc, Sunnyvale, CA), with a 50 microns spot size. The two-views were displayed on a single 21-inch, 2560x2048 gray scale monitor (ORWIN Associates, Amityville, NY), interfaced to a Gateway GP6-266 computer (Gateway, North Sioux City, SD) running Windows 95 (Microsoft

In *Medical Imaging 2000: Image Perception and Performance*, Elizabeth A. Krupinski, Editor, Proceedings of SPIE Vol. 3981 (2000) • 1605-7422/00/$15.00

Corporation, USA). The cranio-caudal view was displayed on the left-hand side of the display, and the medio-lateral oblique view was displayed on the right-hand side.

The observers were instructed to search for malignancy. The experiment was divided into two phases. In the first phase the observers visually searched the images until they felt confident to provide an initial impression about the case, namely, if it was normal or abnormal. During this phase the eye position was tracked using an infrared based system, the 4000SU (Applied Science Laboratories, Bedford, MA). This system has an accuracy of about 1°. Once the observer concluded if it was a normal or abnormal case, they were instructed to pull down a menu on the screen, where they gave their initial impression about the case. This marked the end of the first phase, and the eye-tracking system was turned off for the second phase, in which the observers freely used a digital zoom window to further study any areas of the image where they suspected that a malignant lesion was visible. This zoom window was about 401 x 401 pixels wide, and it was centered at the location indicated by the observer using a mouse-controlled cursor. Inside the zoom window the image was seen at its original resolution of 50 microns. In this phase they were instructed to, upon detecting a malignant lesion, place a mouse-controlled cursor over the center of the lesion and click. This action would prompt a menu to appear in the screen, where they answered what type of abnormality they had found (mass, calcification, architectural distortion) and how confident they were that it was indeed malignant (low, medium and high confidence). These responses were saved to a file that also contained information regarding the time when the decision was made.

Unbeknownst to the observers, the locations and the duration of the zooms, as well as their sequence, were also recorded to a file. This allowed us to keep track of the areas that attracted the observers' attention, and also how conspicuous a stimulus element had to be in order to be zoomed (that is, were the most conspicuous elements zoomed in first?).

Based upon knowledge provided from pathology reports and posterior films, where the cancer was reported, one of the authors (SPW) marked the coordinates of all of the lesions present in this test set and determined their nature (mass, calcification, architectural distortion). This data allowed us to build a truth table, against which we compared the observers' assessment, and rated their decision outcomes as being True Positives (TPs), False Positives (FPs), True Negatives (TNs) and False Negatives (FNs).

3. Results

In this section we will present the results of using the digital zoom window, and compare these with data generated by the eye-tracking system in the first phase of the experiment. Eye fixations were clustered by grouping raw points of eye position data using certain rules. For example, to be included in a cluster the points had to occur in sequence and in the same neighborhood. Furthermore, the fixations had to fall within a grouping that did not exceed 2.5 degrees. If the distance was greater than that, a new cluster was created, having as center the centroid of that group of fixations. For each cluster, the dwell time on the location of the cluster was calculated by multiplying the total number of data points inside that cluster by 1/60, which is the sampling rate for the ASL system.

3.1. Comparing the use of the zoom window with the clusters

In order to compare the two measures of observers' attention, namely, the use of the zoom window and the clustering of eye position, we calculated the mean number of zooms and clusters per case type. Furthermore, we also measured the percentage of clusters (> 1000ms) that were later zoomed, as well as the percentage of zooms that occurred in locations where a cluster (> 1000ms) existed during the scanning phase. This is shown on Table 1.

Case Type	Mean # of Clusters	Mean # of Zooms	% of Clusters that were later zoomed	% of Zooms that occurred in locations of clusters
R	10.593	6.767	30	76
N	7.680	3.000	26	86
NR	9.748	5.673	29	73

Table 1. Average use of the zoom window in comparison with the visual dwell clusters

3.2. Comparing zoom and dwell times during search with the responses made

Table 2 relates the decisions made by the observers, per image type, with the locations that attracted their visual attention during phase 1 and the ones in which they zoomed on phase 2. As it is clear from this table, most of the locations that elicited a response from the observers were either zoomed or received significant (>= 1000 ms) visual dwell during the scanning phase. Furthermore, the differences in the percentages of the decisions that attracted visual attention during the phase 1 from the ones that were zoomed on phase 2 was not statistically significant.

Image Type	Decision Outcome	Location received long (>1000ms) visual dwell	Location was zoomed on Phase 2
R	FN	29%	58%
	FP	42%	77%
	TP	71%	89%
NR	FN	46%	56%
	FP	67%	91%
	TP	54%	91%
N	FP	25%	88%

Table 2. Relationship between the areas in the images that yield a decision outcome and the percentage of them that were either looked at, during the first phase, or zoomed in, during the second phase of the experiment.

3.3. Effect of zoom order on decision type

In order to assess if the most conspicuous elements were zoomed in early or late during the zooming phase, we have numbered the zooms according with the order in which they occurred, and we have related this order to decision outcome. This is shown in Table 3.

Case Type	Zoom Number	Decision Outcome	Statistically Significant Difference Yielded
R	1st	TP	Between FN and TP (Scheffe's test, $p < 0.05$)
	2nd	TP	
	3rd	FP	Between FP and TP ($p < 0.05$)
	4th	FN	
N	1st	TN	There were no statistically significant differences
	2nd	FP	
NR	1st	TP	Between FN and TP ($p < 0.05$)
	2nd	FP	Between FP and TP ($p < 0.05$)
	3rd	FN	

Table 3. Relationship between order in which the zoom occurred and the decision outcome that it yielded, as well as the statistically significant differences between the decision outcomes., as measured by zoom order.

3.4. Effect of zoom length, per zoom number, on decision type

Considering that the order in which the zooms occurred was directly related to the decision outcomes, we decided to verify if the duration of the zoom, measured by how long the observer kept the zoom window fixed in one location, had any significant correlation with the decision outcomes.

For the R cases, for the eighth zoom, there were statistically significant differences between FNs and FPs ($p < 0.05$) and between FPs and TPs ($p < 0.05$).

For the N cases, for the first zoom, it lasted about 4 seconds when it yielded a TN decision, whereas it lasted about 9 seconds when it yielded a FP. This difference was statistically significant ($p < 0.05$).

For the NR cases, for the second zoom, there were statistically significant differences between FNs and FPs. In this case the FNs lasted about 3 seconds, whereas the FPs lasted about 9 seconds.

When comparing these numbers with those yielded by the visual dwell, during the phase 1, on the locations where later the observers indicated (or fail to do so) the presence of a malignant lesion, there were no statistically significant differences for the R and N cases. For the NR cases, there was a statistically significant difference was between FNs and FPs ($p < 0.05$). In this case the dwell on the FNs lasted 1310ms, whereas the dwell on the locations of the FPs was 1970ms.

3.5. *Effect of zoom on performance*

In order to assess if the use of the digital zoom window helped or hurt performance, we have used the first impression provided by the observers, as well as the locations of the clusters with a long visual dwell (>= 1000ms) to score the observers' performance before they were allowed to zoom in on the regions of interest. Thus, for example, if on a lesion-free image the observer had 3 clusters of significant visual dwell, but called the case 'normal' at the end of phase 1, then we scored the observer as having made 3 TNs. On the other hand, if the observer called the same case 'abnormal' at the end of his or her run, then we scored the observer as having made 3 FPs on that image. The same reasoning follows on cases that contained a lesion. Because there was no information available, on phase 1, about the confidence of the observers on their decisions, then ROC analysis could not be used, and we have scored the observers' performance using log odds. Table 4 lists the values for before and after zooming was allowed.

	Initial Impression	After Zooming
NR	Lo = -0.40	Lo = -0.33
R	Lo = -0.36	Lo = 0.88

Table 4. Log odds for the observers performance before and after zooming was allowed.

It can be shown that the gain in the True Positives was about 64% for the R cases, but for the NR cases there was an actual loss of 19% in performance, meaning that the False Positives overtook the True Positives once the use of the zoom window was allowed.

4. Discussion

In this paper we have shown that a digital zoom window can be used to monitor the regions in the image that attracted the observers' attention, as opposed to an invasive infrared eye-tracker. The zoom window was used in the locations of the majority of the decisions made by the observers, even the False Positives and the False Negatives. Furthermore, most zooms occurred in locations where the observers had had long (> 1000 ms) visual dwell. Moreover, the order in which the zooms occurred, as well as the length of the zooms, yielded statistically significant information about decision outcome, which makes the use of a digital zoom window an interesting alternative to aid the observers in improving performance when reading a mammogram test set. Zooming significantly improved performance on the R cases; unfortunately it had the apparent effect of raising the noise level in the cases where a subtle cancerous lesion was present, which decreased performance. Because the test set chosen for this experiment was so heavily biased towards subtle lesions, this decrease in performance was significant. It is unclear if in clinical conditions the use of the zoom window could actually help radiologists to make fewer False Positives and, most importantly, fewer False Negatives.

5. Acknowledgements

This work was partially supported by Grant DAMD17-97-1-7103 between the USAMRMC and C. F. Nodine.

6. References

1. C. F. Nodine, H. L. Kundel, S. C. Lauver and L. C. Toto, "Nature of Expertise in Searching Mammograms for Breast Masses", Academic Radiology, 3:1000-1006, 1996.
2. E. A. Krupinski, C. F. Nodine and H. L. Kundel, "Enhancing Recognition of Lesions in Radiographic Images Using Perceptual Feedback", Optical Engineering, 37:813-818, 1998.

4

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Vol. 1, No. 26
ISSN 1605-7422

Reprinted from

Medical Imaging 2000

Image Perception and Performance

16–17 February 2000
San Diego, California

Proceedings of SPIE
Volume 3981

Image Structure and Perceptual Errors in Mammogram Reading: A Pilot Study

Claudia Mello-Thoms①,②, Stanley Dunn②, Calvin F. Nodine① and Harold L. Kundel①
①University of Pennsylvania School of Medicine, Department of Radiology
②Rutgers University, Department of Biomedical Engineering

Abstract

Early detection of breast cancer is very desirable, considering that it can significantly change the prognosis for a woman diagnosed with this disease. Nonetheless 10-30% of all breast cancers are missed by the radiologist, albeit they are visible in the mammogram. In this work we have studied the underlying structure of the image in the location of the lesions that were missed and the ones that were found, as well as in the locations of the lesions that did not exist but were reported by the radiologist. We have shown that there is a statistically significant difference in the information content of different frequency bands that results in various decision types. We have also shown that it is possible to use a pattern classifier, based upon the information contents of the spectral decomposition of a local image region, to predict the most likely decision outcome.

Keywords: Image structure, perceptual errors, mammogram reading, wavelet packets.

1.Introduction

Early detection can significantly change the prognosis for a woman diagnosed with breast cancer. Thus, renewed efforts have been made to develop accurate imaging techniques that can detect abnormalities of smaller sizes. Nonetheless, a problem that is usually overlooked when considering such imaging techniques is the radiologist's ability to correctly interpret what is on the image. It has been shown [1] that 10-30% of all breast cancers are missed, being only found retrospectively, albeit they are visible in the mammogram. Furthermore, from these, 65% are fixated by the high-resolution central/foveal vision [2]. In other words, these cancers are not missed because of search errors, but because of perception and decision-making errors.

Kundel and Nodine [3] have derived a model that links perception and decision making in medical image reading. This model predicts that perception, and ultimately decision making, start out with a global impression of what is in the image. This global impression is compared to a cognitive schema, stored in memory, of similar images that the observer has seen in the past. This comparison flags regions of potential abnormality, which the observer examines by visually scrutinizing the area with the high-resolution fovea. This results in the extraction of features that are processed and used for object categorization. If a positive fit is found with some representation in memory, additional visual search is performed, until an internal threshold is crossed, and the abnormality is decided positive or negative.

Many factors have been shown to play a role in aiding or preventing lesion detection. Among these, the relationship between the abnormality and the background tissue surrounding it has been shown to be one of the most important. Burgess and colleagues [4] have shown that, in mammograms, lesion detectability is not related to the size of the lesion, but rather to a power law which takes into account the signal energy and the background structure power spectrum. This means that even large lesions can be missed, if certain conditions hold between the lesion and its surrounding tissue.

In this paper we will examine the relationship between breast masses and their surrounding tissue as a function of what decision type they yield, namely, if they yield True Positives (that is, the observer correctly finds a malignant abnormality present in the mammogram), True Negatives (if the observer correctly interprets normal tissue as being lesion-free), False Positives (when the observer incorrectly interprets normal tissue as being malignant) and False Negatives (when the observer fails to indicate a malignant lesion that is visible in the mammogram).

2.Materials and Methods

Eight experienced observers (3 mammographers from the staff of the Hospital of the University of Pennsylvania, HUP, and 5 fellows undergoing training at the same institution) read 5 two-view (cranio-cauldal, CC, and medio-lateral-oblique, MLO) mammogram cases. All cases had a malignant mass visible in at least one view. One case contained multiple malignant masses, visible in both views. These cases were obtained from the archives of HUP. The films were digitized using a Lumiscan Model 100 digitizer (Lumisys Inc, Sunnyvale, CA), using a 100 microns spot size. The two-views were displayed

In *Medical Imaging 2000: Image Perception and Performance*, Elizabeth A. Krupinski, Editor, Proceedings of SPIE Vol. 3981 (2000) • 1605-7422/00/$15.00

on a single 19-inch, 2048x2048 gray scale monitor (GMA 201, Tektronix, Beaverton, OR), interfaced to a Sun Sparc computer (Sun Microsystems, Sunnyvale, CA).

The observers were instructed to search for malignancy, and freely examined the cases until they felt confident to point out if and where a malignant lesion was present. The eye position of the observers was monitored during search, and it was used to determine the areas in the image that attracted the observers' attention.

The eye position of each of the observers was played back over the mammogram case examined, and from each case 10 regions were manually extracted using a mouse-controlled cursor. These regions contained true lesions that were indicated by the observer (and were labeled TP), true lesions that were missed by the observer (labeled FN), lesion-free areas that were indicated by the observer as being lesion-containing areas (FP) and lesion-free areas that were correctly interpreted by the observers as being normal tissue (TN).

Each of these regions was processed using a filter bank that contained quadrature-mirror filters, using a process known as Wavelet Packets. This tree was two-levels deep. During the decomposition of each region some statistical parameters were calculated for each frequency band, including the mean and standard deviation and the signal energy in that band. Each band was represented by a combination of two numbers, one that indicated where in the tree the band was located (levelwise) and one that indicated if the signal being processed had been low-(or high-)passed in the previous step and was now being low-(or high-)passed.

3. Results

The mean values for the energy in the different frequency bands is listed in Table 1.

band	Mean energy value		band	Mean energy value
00	11327.45		22	2.08
01	4.29		23	1.99
02	20.17		30	16.11
03	25.01		31	2.08
10	42502.70		32	59.05
11	0.44		33	0.44
12	16.11		40	19.86
13	19.86		41	1.99
20	0.44		42	0.44
21	11.38		43	73.77

Table 1. Mean values for the energy per frequency band.

As shown, there is a wide variability in the information contents of each of the bands. Thus, the bands were divided in three classes: the low energy (which had a mean <= 10), the medium energy (10 < mean <= 50) and the high energy bands (mean > 50). Thus, in order to assess the contribution of each band on the decision outcomes an ANOVA analysis was run. In all Scheffe tests listed below, the significance level was 5%. Table 2 lists these results.

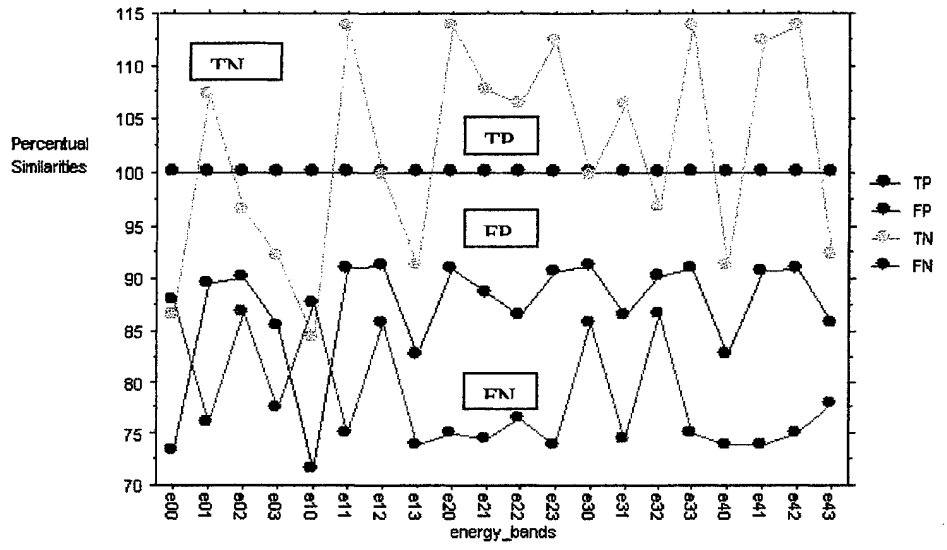
Band class	Band Name	Contributed for the differentiation between:
High	00 10	TPs from FPs ($p < 0.05$) TPs from FPs ($p < 0.05$)
Medium	21	FPs from TNs ($p < 0.05$)
Low	42	FPs from TNs ($p < 0.05$)

Table 2. List of the energy bands that contributed to the differentiation of pairs of decision outcomes, as tested using Scheffe's test.

The importance of these results stems from the fact that they clearly state that there exist differences in the energy contents, per frequency band, of the regions of the image that result in different decision outcomes.

Furthermore, if one considers the mean values of energy on the different frequency bands that lead to True Positive decision outcomes as a base value, then the breakdown of energy, in percentual values, relative to the levels of the TPs, for the remaining decision outcomes is shown in Figure 1.

Figure 1. Percent differences between the TPs and FPs, FNs and TNs.



As Figure 1 shows, the energy contents are generally higher for the True Negatives, particularly for the intermediate energy bands. As their contents begin to change, that is, to lose power in the majority of the bands, the False Positives are formed. As power continues to decrease, the False Negatives come about. An ANOVA was run on these percentual differences, and it was found that there are statistically significant differences between FNs and FPs (Scheffe test, $p < 0.05$), FNs and TNs ($p < 0.05$) and FPs and TNs ($p < 0.05$).

Knowing from the first ANOVA which bands are responsible for the differentiation between different pairs of decision outcomes, we decided to use a Neural Network to predict, from the values of the energy in the high- and intermediate-bands, the decision outcome that that a particular region of image would yield. Additionally a parameter that ranged from 1 to 8 was used to inform the network which observer had provided the data being examined. The reasoning here is that different observers may perceive the same region of the image very differently; for example, a more experienced observer may be able to detect a subtle mass whereas another observer may not see anything.

Using an Adaptive Resonance network the results shown on Table 3 were obtained, in terms of correct and incorrect responses. Once more, the purpose of the network was not to determine if an abnormality was present or not on a particular region of the image, but rather, to determine which decision outcome was more likely for a given observer when examining that region of the image.

Class	Correct Predictions	Incorrect Predictions
TP	44/69 = 64%	25/69 = 36%
FP	46/60 = 77%	14/60 = 23%
TN	37/53 = 70%	16/53 = 30%
FN	5/18 = 28%	13/18 = 72%

Table 3. Percent values for correct and incorrect decision outcomes as predicted by the neural network.

This result clearly indicates that it is possible to separate TPs, TNs and FPs based upon the energy decomposition of the region indicated by the observer. Nonetheless, the results for the False Negatives were not good. This is certainly a reflection of the limited number of such samples that was available in this pilot study.

4.Discussion

These results indicate that there is a particular configuration of energy, in the frequency domain, that leads observers to detect true lesions. Furthermore, there also exist particular energy configurations that will likely lead the observers to make False Positives, False Negatives and True Negatives.

When using a pattern classifier to automatically predict which decision outcome will a particular combination of energy in different frequency bands yield, we found that the TPs, FPs and TNs could be reliably predicted, but, due to the small sample size, the same was not true for the FNs. We believe that as our research proceeds, with a much larger database, the results for the FNs will significantly improve.

5.Acknowledgement

This work was partially supported by Grant DAMD17-97-1-7103 between the USAMRMC and C. F. Nodine.

6.Reference

1. M. Giger and H. MacMahon. Computer-aided diagnosis. Radiologic Clinics of North America, 34:565-596, 1996.

2. E. A. Krupinski and C. F. Nodine. Gaze duration predicts the location of missed lesions in mammography. In A. G. Gale et al., editor, Digital Mammography, Elsevier Science B.V., 1994.

3. H. L. Kundel. Perception and representation of medical images. In Proceedings of the SPIE, Image Processing, vol 1898, pp 2-12, 1993.

4. A. E. Burgess, F. L. Jacobson and P. F. Judy. On the detection of lesions in mammographic structures. In E. Krupinski, editor, Proceedings of the SPIE Conference on Image Perception and Performance, vol 3663, pp 304-315, 1999.

A Perceptually Tempered Display for Digital Mammograms[1]

Harold L. Kundel, MD • Susan P. Weinstein, MD • Emily F. Conant, MD • Lawrence C. Toto, BS • Calvin F. Nodine, PhD

The cathode ray tube of a workstation for use with digital mammograms was calibrated with a photometer to produce an input-output characteristic curve similar to the perceptually linear curve defined by a current display standard. Then, a test pattern consisting of bars of increasing intensity containing disks of decreasing contrast was used by an observer to estimate the minimal detectable contrast (MDC) at different levels of display luminance. The MDC was modeled by a parabola. The shape of the parabola was determined by the observer's perceptual responses, and the range was determined by the maximum and minimum pixel values of the breast parenchyma. As each mammogram was displayed, the contour of the breast was automatically found and pixels within the breast image were sampled to determine the pixel values that were used to compute the maximum and minimum pixel values. The parabola was integrated to determine the look-up table for the initial MDC-tempered display of the mammogram. Preliminary observer performance tests showed no significant differences in the accuracy and speed of three radiologists who read a set of mammograms when the MDC-tempered display was compared with the perceptually linear display.

Abbreviations: CRT = cathode ray tube, MDC = minimal detectable contrast, ROC = receiver operating characteristic

Index terms: Breast, 00.99 • Images, display • Radiography, digital

RadioGraphics 1999; 19:1313–1318

[1]From the Pendergrass Diagnostic Research Laboratory, Department of Radiology, University of Pennsylvania Medical Center, 308 Stemmler Hall, 3600 Hamilton Walk, Philadelphia, PA 19104. Recipient of a Certificate of Merit award for an *info*RAD exhibit at the 1998 RSNA scientific assembly. Received April 9, 1999; revision requested May 12 and received June 10; accepted June 21. Supported by grants DAMD17-96-1-6153 and DAMD17-97-1-7130 from the U.S. Army Medical Research and Materiel Command. **Address reprint requests to** H.L.K.

■ INTRODUCTION

Given the present state of the art, a static cathode ray tube (CRT) display can simulate but not duplicate the image quality of a film mammogram displayed on a light box. The film is displayed at higher luminance and has greater spatial resolution and a wider gray-scale range (1). However, the film captures and displays the image by using a fixed set of predetermined display parameters. The CRT display can be adjusted to explore the full range of contrast and resolution available in the digital image by using the window level to change the gray-scale range and zoom-rove functions to change spatial resolution. The appearance of the gray scale within the image can also be changed by modifying the input-output transfer characteristic of the CRT by using look-up tables. The overall appearance of the image can also be changed in more fundamental ways by the application of image processing such as edge enhancement. In this article, we consider only the effects of modifying the input-output transfer characteristic. To have identical images look alike when displayed on different CRTs, a display standard called perceptual linearization has been proposed (2,3). When the standard is used, equal changes in the pixel gray-scale value produce equal changes in the just noticeable difference (JND) of luminance in the image.

A display standard provides an equivalent starting place for each image but may not provide the best distribution of gray levels for a particular image in a particular reading environment. For example, the image may be too dark or too light, just as an image on film may be under- or overexposed. The ability of the human eye to see the intensity difference between two areas in an image (contrast sensitivity) depends on the average intensity of the light reaching the eye (4). The average intensity of the light reaching the eye is termed the *adapting luminance*. When the adapting luminance is very different from the average luminance of the area of interest in the image, the ability to see contrast is decreased. This is the reason why masking the bright areas on a film illuminator improves the appearance of images, particularly dark ones. Most of the light that affects contrast sensitivity comes from the displayed image, but some comes from room illumination including that which is reflected from the CRT surface. Once the room illumination has been minimized, the contrast sensitivity of the eye can be maximized by adjusting the gray scale to smooth out extreme variations in brightness within the image (5,6).

Using a model proposed by Mokrane (7), Liu and Nodine (8) developed an algorithm that equalizes perceived contrast over the image, with some starting level of adapting luminance assumed. Contrast in the image is modified on the basis of the theoretical threshold-contrast curves of Heinemann (4). The workstation described herein extends the work of Liu and Nodine (8) to include adjustment of the input-output transfer characteristic for ambient illumination and for the gray-scale range of the particular mammogram being displayed. In this article, we describe the display station, development of the perceptually tempered display, and evaluation of the display station.

■ THE DISPLAY STATION

The display station shown in Figure 1 uses a computer (model GP6-266; Gateway 2000, Sioux City, Iowa) with a Pentium II processor (Intel, Santa Clara, Calif). The computer is interfaced to a gray-scale monitor (model DS5000L; Orwin Associates, Amityville, NY) by means of an interface board (model Md5/PCI-1; Dome Imaging Associates, Waltham, Mass). The computer software is written in IDL (Research Systems, Boulder, Colo), a high-level graphics language.

Before use of the display station, the video monitor was photometrically calibrated. A photometer (model J17; Tektronix, Beaverton, Ore) interfaced to the computer was used to measure the intensity of a 10 × 10-cm square of uniform luminance located in the center of the display surface. (The luminance of a display such as a CRT or a film illuminator is measured in foot-lamberts or candelas [cd] per square meter [1 foot-lambert = 3.4 cd/m^2].) The intensity of the display surrounding the square was set at a luminance of 55 cd/m^2, which was produced by a pixel driving intensity value of 128. The luminance was measured over 17 equally spaced pixel driving intensity values from 0 (black) to 255 (white); these pixel driving intensity values corresponded to a luminance of 1.7–343 cd/m^2. Digitization and logarithmic transformation of the photometric data were performed; they were then displayed on the CRT along with an ideal perceptually linearized curve. The brightness and contrast controls were adjusted until the calibrated curve visually matched the ideal curve.

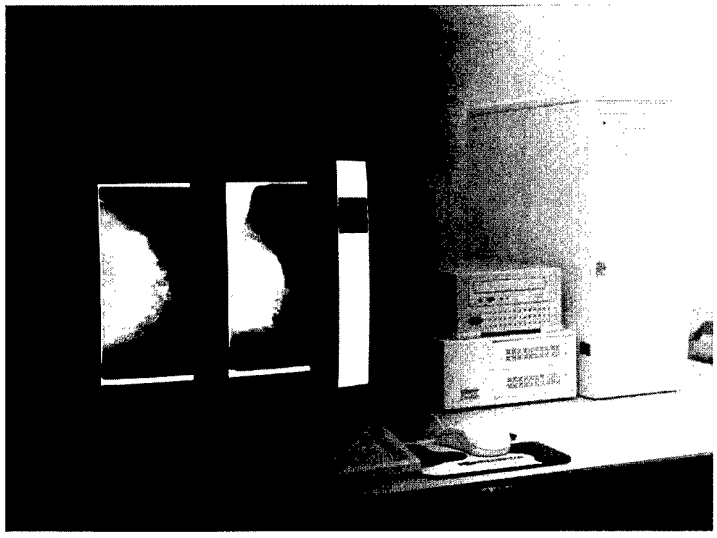


Figure 1. The digital mammography workstation.

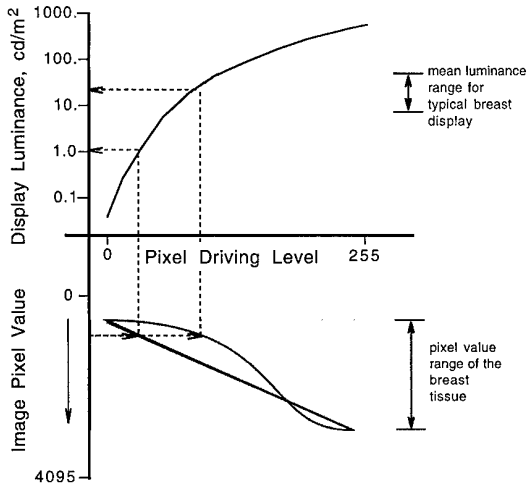


Figure 2. Input-output transfer characteristic of the CRT (top curve) and final minimal detectable contrast (MDC) look-up table (bottom curve). The curves have a common pixel driving level axis. The nonlinearity of the MDC curve is exaggerated for purposes of illustration; the actual difference from the linear curve is usually more subtle. The effect of the MDC look-up table on the displayed image can be seen by following the dotted lines, which represent extrapolation from the image pixel value to the display luminance.

Once the CRT is calibrated, it needs only occasional adjustment. The shape of the input-output transfer characteristic adjusted according to the perceptually linear display standard is shown in Figure 2 (top curve).

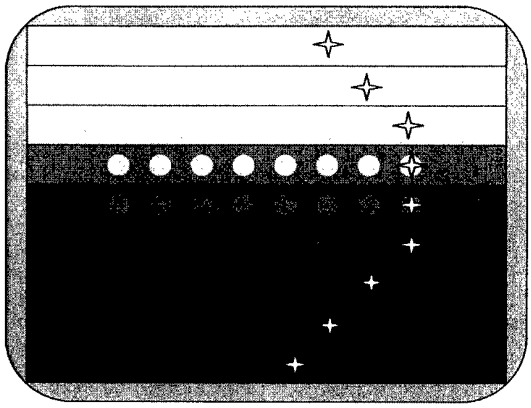


Figure 3. MDC test pattern. ✧ = typical observer response.

■ DEVELOPMENT OF THE PERCEPTUALLY TEMPERED DISPLAY

• Estimation of the MDC

The MDC test pattern consists of nine horizontal bands of increasing intensity (Fig 3). Each band contains eight disks of decreasing contrast. This test pattern was displayed for each observer prior to a viewing session. The observer's task was to choose the "least visible" disk in each band. The observer's responses are affected by the display contrast and the ambient room lighting. A parabola was fitted to the contrast of each indicated disk and the intensity of

the horizontal band; this parabola approximates the dependence of the observer's contrast sensitivity on display luminance at the level of ambient illumination (Fig 4).

• Determination of the Range of Pixel Intensities of the Mammogram

As each case is displayed, the maximum and minimum pixel intensity in the breast parenchyma is determined by sampling over a region that includes breast tissue out to just beyond the skin line, thus excluding the extremes of pixel driving levels due to lead markers, labels, and cassette edge artifacts. Determination of the pixel intensity range is performed with a boundary detection procedure: After applying a median filter, an intensity threshold value 5% above the background (dark level) is selected. By means of this threshold, the breast image is transformed into a binary image and a contour is determined on the resultant image. Image intensities are then sampled on the original breast image along 30 equally spaced lines (Fig 5).

• Production of the MDC Look-up Table

The best-fit parabola for MDC versus displayed luminance is integrated to produce an MDC-corrected look-up table. The maximum and minimum pixel driving levels determined from the mammogram are applied to the MDC-corrected look-up table so that the output intensity just matches the input intensity (Fig 2 [bottom half]). The MDC look-up table is designed to equalize the detectability of equal-contrast targets regardless of the regional mean pixel intensity surrounding the targets. The advantage of redistributing the contrast in this "tempered " fashion is to provide an initial view that allows visual access to the dark regions (fat, skin line) as well as the light regions (muscle, fibrous and ductal tissue). The viewers are still able to manipulate the gray scale of the image. All of the calculations and look-up table manipulations are done by using a 12-bit pixel intensity scale. This scale is transformed into an 8-bit scale for display.

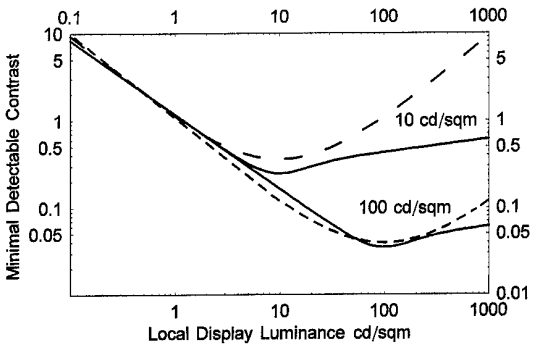
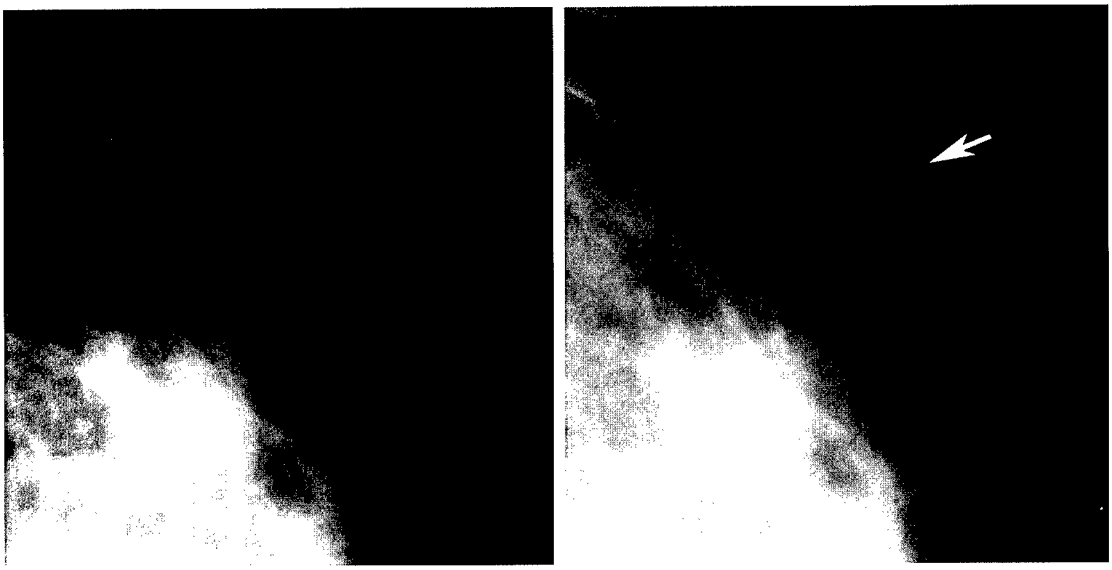


Figure 4. Approximation of the contrast sensitivity curve with a parabola. Heinemann (4) measured human contrast sensitivity at different levels of adapting luminance. Examples of this relationship at two adapting luminance levels are shown (solid lines). In reality, there is a whole family of curves of similar shape that have a minimum that shifts with the adapting luminance. Consider the lower curve, which corresponds to an adapting luminance of 100 cd/m^2. The eye is maximally sensitive at a display luminance of 100 cd/m^2, with an MDC of about 0.05. However, an object located in a dark part of the image at 10 cd/m^2 would have to have a contrast of 0.1 to be seen. The practical solution in radiology is to use a spotlight to raise the luminance to 100 cd/m^2 and improve the contrast sensitivity. As the adapting luminance decreases, the curves shift upward and maintain roughly the same shape. Attempts have been made to fit the curves from Heinemann's experimental data with simple equations (5). The algorithm of Liu and Nodine (8) required advanced information about adaptation level and was computationally intensive. We simplified that algorithm by assuming that a parabola (dashed lines) could be used to approximate contrast sensitivity at different levels of adapting luminance. The fit is reasonable at high adapting luminance (100 cd/m^2), where radiologists prefer to operate. The fit for a lower adapting luminance (10 cd/m^2 [upper curve]) is not very good. However, this luminance is well below a practical average viewing luminance.

• Display of the Images

The CRT is photometrically calibrated as part of the regular quality assurance program. The MDC calibration is performed before each reading session with the ambient illumination set at 1.6 lux at the location of the observer's eyes. The calibration takes approximately 15–20 seconds to complete. The correction of each image is done off-line prior to the test. Observers are able to use a single slider to adjust the MDC look-up table. The slider can smoothly adjust the dis-



a. b.

Figure 6. Mammographic image displayed with standard perceptually linearized display **(a)** and MDC-tempered display **(b)**. The skin line (arrow in **b**) is not visible in the standard perceptually linearized display **(a)**.

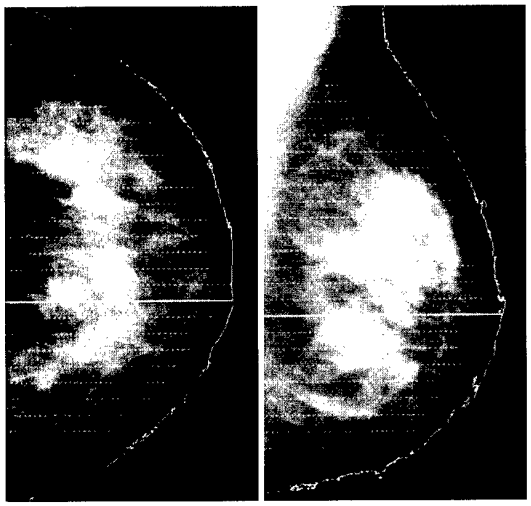


Figure 5. Pattern used for sampling pixel intensities on the breast images. The intensities of the breast are sampled, and nontissue regions beyond the breast are eliminated.

play from a look-up table, which produces the baseline perceptually linearized display standard, up to a maximum MDC setting. Figure 6a shows a breast image displayed with standard perceptually linearized display; Figure 6b shows the image displayed with MDC-tempered display, which allows visualization of the skin line.

■ EVALUATION OF THE DISPLAY STATION

Our development cycle includes periodic benchmark testing by using a sample of cases from a digital database of normal and abnormal mammograms, in which all of the malignancies and many of the benign lesions are histologically proved. The mammograms were originally obtained on film and were digitized to a pixel size of 100 μm with a digitizer (Lumiscan 100; Lumisys, Sunnyvale, Calif). Readers are shown a craniocaudal view and a mediolateral oblique view and are asked to move a pointer on the display to any potential malignant lesion and click the mouse. The response time from the start of viewing each case and the location of the pointer are recorded by the software. After the click, a pull-down menu appears; the reader must select one or more diagnoses (ie, mass, calcification, or architectural distortion) and indicate a confidence level for malignancy. These data are used to compute a receiver operating characteristic (ROC) curve and determine the area under the curve.

Two mammographers (S.P.W., E.F.C.) and a general radiologist (H.L.K.) were tested on 75 mammograms: 25 with malignancies, 25 with

Table 1
Areas under the ROC Curve for Perceptually Linear Display versus MDC-tempered Display

Reader	Linear Display	Tempered Display	Difference
1	0.910	0.930	0.020
2	0.861	0.869	0.008
3	0.627	0.750	0.123
Mean	0.799	0.850	0.050*

*Standard deviation = 0.063.

Table 2
Time to First Decision in Seconds for Perceptually Linear Display versus MDC-tempered Display

Reader	Linear Display	Tempered Display	Difference
1	76	51	−25
2	55	84	29
3	51	47	−4
Mean	61	61	0*

*Standard deviation = 27.

benign lesions, and 25 that were normal. Table 1 is a comparison of the areas under the ROC curve. Although each reader did better with the MDC-tempered display, the difference was not significant when tested with a paired *t* test. The time to the first pointing out of a lesion was very variable but on average was not different for the two display modes (Table 2).

■ CONCLUSIONS

The accuracy and speed of the tempered display function are equal to those of the standard perceptually linearized display function when used on a moderately bright monitor (300 cd/m^2). With the tempered display function, the initial view of the image provides visual access to lighter and darker regions of display with some sacrifice of visual access to medium-intensity regions. The display can be adjusted by moving a single slider, which is an attempt to simplify the user interface. Development of the display station is continuing with the addition of the use of verbal commands to modify display parameters and an eye position–contingent roving window.

■ REFERENCES

1. Blume H, Roehrig H, Browne M, Ji TL. Comparison of the physical performance of high resolution CRT displays and films recorded by laser image printers and displayed on light-boxes and the need for a display standard. Proc SPIE 1990; 1232:97–114.
2. Johnston RE, Zimmerman JB, Rodgers DC, Pizer SM. Perceptual standardization. Proc SPIE 1985; 536:44–49.
3. Blume H, Hemminger BM. Image presentation in digital radiology: perspectives on the emerging DICOM display function standard and its application. RadioGraphics 1997; 17:769–777.
4. Heinemann E. The relation of apparent brightness to the threshold for differences in luminance. J Exp Psychol 1961; 61:389–399.
5. Cobra D. Image histogram modification based on a new model of visual system nonlinearity. J Electron Imaging 1998; 7:807–815.
6. Pizer SM, Amburn EP, Austin JD, et al. Adaptive histogram equalization and its variations. Comput Vision Graph Image Process 1987; 39:355–368.
7. Mokrane A. A new image contrast enhancement technique based on a contrast discrimination model. Comput Vision Graph Image Process 1992; 54:171–180.
8. Liu H, Nodine CF. A generalized image contrast enhancement technique based on the Heinemann contrast discrimination model. J Electron Imaging 1996; 5:388–395.

CHAPTER 19
The Nature of Expertise in Radiology

Calvin F. Nodine, Claudia Mello-Thoms
University of Pennsylvania Medical Center

CONTENTS

19.1 Introduction

This chapter is about expertise in radiology. In the domain of radiology, expertise is largely acquired through massive amounts of case-reading experience. But just as everyone who is taught how to read is not an expert reader, so too everyone who is taught how to read medical images is not an expert image interpreter. The criterion that defines an expert medical-image interpreter is consistent and reliably accurate diagnostic performance. Nothing less will do. For example, despite intensive study and training, it has been shown that radiology residents at the end of residency training are significantly below the average of a large national sample of U.S. radiologists in overall accuracy of screening mammograms for breast cancer (Nodine, Kundel, Mello-Thoms *et al.*, 1999). This finding is not surprising when considered within the framework of research on expertise, which stresses that expert performance in many domains is, statistically speaking, rare, and usually accomplished only after extensive training and practice (Chi, Glasser, Farr, 1988; Ericsson and Charness, 1994).

We view expertise as the ability to acquire and use contextual knowledge that differentiates one from one's peers in a particular field. In this sense, expertise is a contextual concept, because the knowledge-structured skills that make an expert in one domain do not transfer to other domains (Nodine and Krupinski, 1998; Patel and Groen, 1991). Moreover, expertise is composed of a sum of different parts, each having a unique influence on the total. For example, in the context of medical image interpretation, an expert is someone that has had more experience, meaning diagnosed more cases, thus providing a broader range of variations of normalcy against which to differentiate abnormal findings. An expert is also someone who has a natural talent to perform within a chosen domain. Again, from a radiological perspective, different radiologists may have seen a similar number of medical images, but some will stand out in their ability to diagnose abnormalities, and perform the task faster. This component of expertise is called by us talent, but there is no doubt that motivational factors may be coloring what is termed talent (Ericsson, 1996, p. 27; Ericsson and Charness, 1994, pp. 728–729).

Although expertise has been extensively studied in many domains, the concept is still very elusive. If at this point one was able to pinpoint what makes an expert in any given field, one could certainly go out and create an artificial expert in that field simply by teaching a machine the skills that make one an expert. This has been tried many times, and some success has been achieved. Expert systems have been developed to find calcifications in mammograms (Nishikawa, Jiang, Giger *et al.*, 1994), to detect signs of lung cancers in chest radiographs (Lo, Lin, Freeman *et al.*, 1998), to differentiate benign from malignant lesions in mammograms (Zheng, Greenleaf, Gisvold, 1997). However, we are still very far from having an intelligent system that can actually read and interpret a medical image as reliably, accurately, and efficiently as a human expert.

The reason for this may be in the nature of expertise itself. As previously mentioned, medical expertise is formed by two parts, one that is computable, which responds to training by learning, and one that is uncomputable, which is independent

of training, referred to as talent. We can design models that approximate the logical reasoning of experts when they are examining an image and making a decision, but there has been little success modeling internal processes that are responsible for the talent part (see Ericsson and Charness, 1994). Furthermore, these processes do not seem to arise from a structured thinking hierarchy, but rather seem to evolve spontaneously.

Thus, one is forced to consider the possibility that machine expertise will be restricted to the acquired part that makes up human expertise, which is related to training, to structured knowledge, to rule-based thinking. This is not to say that the performance of expert systems should not be compared with human experts, but rather, that expert systems possess a different kind of expertise. This by no means invalidates the need for intelligent systems in medical diagnosis. As shown elsewhere (Nodine, Kundel, Mello-Thoms *et al.*, 1999) it takes a great number of cases for one to become an expert mammogram reader, and it is here that intelligent systems will find their niche, by either providing a second, informed diagnosis, or by working as tutors, helping less experienced radiologists or radiology residents make as many correct decisions as possible, while keeping errors to a minimum. It is our belief that in contexts where both parts of expertise are operating, expert systems will surpass human performance in the computable part, but remain void when it comes to the talent component of expertise.

19.2 Plan of the chapter

Radiology is largely a visual discipline. This means that rather than relying on direct observation of patients, radiologists rely on interpreting image representations (usually generated by x-ray imaging) to gather diagnostic information about the medical status of patients. They may also read the patient's clinical history either to guide or to clarify image interpretation.

The interpretation of medical images depends on both image perception and cognitive processes. Often-cited perceptual skills include visual search, visual information processing, and visual discrimination and differentiation which are part of perceptual learning. In addition to perceptual skills, the interpretation of medical images depends on cognitive skills primarily related to diagnostic reasoning and decision making. Expertise represents a honing of these perceptual and cognitive skills. But, how much of expertise in radiology is learning to understand what one is looking at anatomically (basic science) and how much is what one sees within a clinical context as signaling pathology (clinical problem solving) is difficult to estimate.

In this chapter we shall summarize some of the findings on expertise in radiology. The theme is to show how image perception interacts with decision making to produce skilled diagnostic interpretation of medical images. The basic information-processing flow to achieve this is: VISUAL SEARCH—OBJECT RECOGNITION— DECISION MAKING. This is our "brand" of expertise theory (Nodine and Kundel, 1987). It is biased toward the perceptual side, whereas

many radiology expertise studies are biased toward the cognitive side. Our perceptual bias is reflected in our choice of theory and methodology. This is true of the cognitive camp as well. Thus, we look at radiology expertise as primarily visual problem solving. Our methods depend heavily on generating theoretical inferences from eye-position data, whereas many studies using the cognitive approach depend on generating theoretical inferences from verbal-protocol data (e.g., Lesgold, Feltovitch, Glaser *et al.*, 1981; 1988). Note that a third approach exists, namely, the connectionist approach (Dawson, 1998), which models information processing in artificial intelligence (AI) using artificial neural networks (or ANNs). This approach will be discussed later in this chapter.

We will use a recent study of mammography expertise to illustrate some basic points: First, how experience influences the acquisition of expertise, and a discussion of the imperfect translation of experience as an error-correction feedback mechanism for training radiology residents. Second, how the three components of information processing, search, recognition and decision making, combine to produce diagnostic-decision outcomes. Third, how the information-processing model works across radiology subdomains by comparing research findings in chest and breast radiology, and pointing out some important differences in the two subdomains that may result in negative transfer.

19.3 Expertise roots

Expertise research has its roots in the intersection of cognitive psychology and computer science, now known as artificial intelligence, or AI. The cognitive psychology side of this research was concerned with identifying human information-processing skills associated with solving intellectual problems (e.g., playing chess, solving physics problems, diagnosing disease), and the computer-science side was interested in modeling cognitive processes by developing programmable algorithms that would generate performance outcomes with the ultimate goal of creating expert systems. The overarching framework for research on expert systems was learning theory generally and problem solving specifically. Man was conceived of as a processor of information, and the process of seeking information was analyzable in terms of contingencies of reinforcement, that is, feedback, that corrected erroneous behavior and thus guided the course of learning.

A lot of water has passed over that dam since AI began the study of expertise. The late Alan Newell, one of the founding fathers of AI predicted in 1973 that "... when we arrive in 1992 (Newell's retirement date from Carnegie-Mellon University) we will have homed in on the essential structure of mind" (Newell, 1973, p. 306). Needless to say, that prediction was a bit optimistic, but it does reflect the enthusiasm and hopes that one of its founders had for AI. Some would say the defining moment for the beginning of AI was George Miller's article on the limits of human information-processing capacity (Miller, 1956). Cognitive psychology was thought, at the time of its inception in the 50's, as reflecting a shift away from behaviorist learning theory toward finding mental structures (rules for learning and

problem solving). Looking back today, almost 50 years later, it is somewhat amusing to see how stubbornly reinforcement (albeit redefined), the cornerstone of behavioristic theory, continues to survive within mainstream learning theory in spite of the cognitive revolution which spured cognitive theory, cognitive science, and artificial neural networks.

As already indicated, research on expertise has been wide ranging and it is not the purpose of this chapter to review it all. Rather, the goal is to provide a glimpse of expertise research by focusing on radiology. When it comes to studying expertise, the domain of radiology is broad. In today's era of specialization in medicine, we have experts in subdomains of radiology, as for example, breast imaging (mammography), thoracic imaging, angiography, etc., and an expert in mammography is unlikely to also be an expert in another subdomain.

The hierarchial ordering from general to specific domains is important to recognize in studying expertise in radiology because, although radiology resident training provides mentoring experiences in a number of subdomains of radiology, the ultimate goal of such training is to make a radiologist who is Master of one subdomain, rather than Jack of all subdomains. The result is that radiology expertise is subdomain specific. This emphasis on subdomain-specific training and experience fine tunes the radiologists such that performance of an experienced mammographer may suffer if asked to interpret a chest radiograph, or performance of an experienced chest radiologist may falter if asked to interpret a mammogram. From the standpoint of a learning theory framework, expertise skills are specific to a given subdomain and do not effectively transfer to other subdomains within radiology. This is true of medical expertise in general (Patel and Groen, 1991).

19.4 Expertise, acquired or innate?

Most expert performance is acquired, not innate (Ericsson and Charness, 1994). This is not to say that native talent plays no role, but its role is limited, particularly in medicine. Perceptual tests to identify visual skills of prospective radiologists have not generally been successful (Bass and Chiles, 1990; Smoker, Berbaum, Luebke *et al.*, 1984). This is because what the test purports to measure (e.g., spatial relations) is either too abstract or too far removed from the radiology task. Thus there is little or no transfer of test skills to radiology reading skills. A good example is finding NINA in Al Hirschfeld's drawings of theater scenes. Hirschfeld's task calls on visual-search skills for locating NINA targets within the theater scene, and object-recognition skills for disembedding letters from features of theatrical scenery in order to recognize NINA's name. These search and recognition skills seem to be very close to what is required of the radiologist searching a chest image for a lung nodule, but we have found that radiologists are no better than laypersons at finding NINAs (Nodine and Krupinski, 1998). This result seems to argue that expertise in radiology is very narrow and subdomaine specific. In the following section we will examine expertise in different domains, and see how expertise in chess and in the medical field compare to expertise in radiology.

19.4.1 Chess-playing expertise

Parallels have been made between the chess master and the radiology expert performing their tasks (Wood, 1999). Both have built extensive, organized, and searchable mental arrays containing task-specific information such as configurations of chess pieces mapped to feasible game-playing moves, or radiographic patterns of anatomy mapped to pathological signs that are used in solving their respective problems. These mental arrays are often referred to as schemas.

Expertise research starting with de Groot's (1965) and Chase and Simons' (1973) seminal studies and generalizing across a number of expertise domains has found practice to be the major independent variable in chess skill. For example, Charness, Krampe and Mayr (1996) have shown that "deliberate practice" is critical for acquiring skill in chess playing. What is meant by deliberate practice is self-motivated effortful study. We believe that this definition, broadly stated, describes the type of study medical residents go through during residency training. Their study is closely supervised by mentors who motivate learning and guide training by drawing on a vast data base of clinical experience. In chess, for example, Charness *et al.* estimate that 32,000 hours of deliberate practice over 9 to 10 years are necessary on average to achieve grandmaster levels (2500 Elo points; Elo was the name of the man who developed the scale). It is important to note the Charness *et al.* distinction between deliberate practice, and casual practice which involves playing games with others. Deliberate practice correlates higher with chess skill ($r = 0.60$) than does casual practice ($r = 0.35$), which lead them to conclude that deliberate practice is the primary change agent influencing chess skill. One reason this conclusion is so important to our discussion of radiology expertise is because acquiring radiology skill depends on highly motivated and supervised learning, and Charness *et al.* have shown that this type of learning produces more effective cognitive-skill outcomes than casual learning and book reading. This realization has important implications for training radiology residents where supervised learning takes the form of mentor-guided experiences.

Another reason to look carefully at chess expertise is because chess skill has a perceptual component of search that draws on schematic representations of chess-move patterns leading to "best" game moves (Gobet and Simon, 1996) in much the same way as radiology skill has a perceptual component of search that draws on schematic representations of normal anatomic patterns against which to compare new image input for signs of abnormality. Studies of expertise in both chess and radiology have been modeled as problem-solving tasks of the general form: SEARCH & DETECT—EVALUATE—DECIDE. Both chess and radiology have a strong visual-spatial component.

19.4.2 Medical expertise

The study of medical expertise draws heavily on linguistic analysis of the semantic content of propositional statements of physicians recorded as verbal protocols, or thinking out loud. The verbal protocols are scored for the recall of

medical-knowledge representations and reasoning processes used to generate either data-driven inferences or hypothesis-driven facts leading to diagnostic explanations (Patel and Groen, 1991). In contrast, the study of radiology expertise, because the focus is shifted from observations of a live patient to medical images, focuses on perceptual analysis of image features, or statements about the interpretation of image features (see Raufaste, Eyrolle, Marine, 1998).

19.4.3 Radiology expertise

The study of expertise in radiology has been limited to the subdomains of chest radiology and mammography. The radiologist's task in both cases has been modeled within a visual problem-solving framework. However, different experimental methods have been used to gain insights into the nature of perceptual-cognitive skills underlying medical-image interpretation. The cognitive approach as exemplified by Lesgold (1984) and Lesgold, Feltovitch, Glaser *et al.* (1981, 1988) uses a form of verbal-protocol analysis involving analyses of observers' diagnostic reports and sketches of abnormal regions to identify cognitive structures that presumably interact recursively between hypotheses and image features to generate diagnostic outcomes.

The perceptual approach as exemplified by our research (e.g., Nodine, Kundel, Mello-Thoms *et al.*, 1999; Kundel and Nodine, 1975) has used a combination of measures derived primarily from eye-position data to characterize schema-driven search strategies leading (or following) focal analyses of perceptual features from which diagnostic decisions are inferred. The eye-position data include percent coverage of the image by a fixed circular field approximating the size of the fovea plus error tolerance, time spent dwelling on selected image location (target or nontarget) referred to as cumulative gaze duration or visual dwell, and time to detect a target referred to as search time to a hit. In our most recent work in mammography, we compared speed-accuracy performance as a function of level of expertise using a chronometric analysis of decision time. Decision time is similar to reaction time used by Posner (1986). We used decision time to measure how training influenced information-processing skills in screening mammograms for breast cancer.

Expertise in radiology, and in medicine more generally, refers to reliably accurate diagnostic problem solving. This does not mean that radiologists are infallible. They make errors, and much of our research and that of others has focused on the error side of the coin rather than the accuracy side (e.g., Kundel, Nodine, Krupinski, 1990; Parasuraman, 1986).

19.4.4 Mammography expertise

To illustrate the importance of training and experience (practice) on the acquisition of medical-image interpretive skills, we recently concluded a study of mammography expertise in which we compared performance of experienced breast imagers (mammographers) with radiology residents undergoing mammography training and mammography technologists. Our study was designed to compare observers having differing degrees of interpretive skill reading mammograms in an

effort to shed light on how such skill is acquired and reflected in performance. As part of this study, in order to provide a clearer picture of how the three groups differ in experience interpreting mammograms, we obtained data about the number of mammographic reports generated by residents and mammographers. This was done as a way of quantifying the amount of mentor-guided image-reading experience residents received. The 19 radiology residents who were part of the study represented mainly third-year ($n = 7$) and fourth-year ($n = 8$) residents, plus 4 fellows, with mammography reading experience varying from 10 to 2465 cases over a 3-year interval. The average reading experience at the end of resident training was 650 cases for our resident sample. Over the same period, each of the 3 mammographers read 9459 to 12,145 cases.

Figure 19.1 shows the relationship between log (base 10) cumulative number of mammogram cases read over a 3-year interval and A1, the area under the AFROC (alternative free operating characteristic) curve, which is a measure of overall diagnostic accuracy.

This figure shows a significant linear-regression fit of the data ($R^2 = 0.667$) with a positive slope suggesting that interpretive-skill competence as reflected by A1 performance (area under the AFROC) increases directly with log case-reading experience. This finding is strikingly similar to that found by Charness *et al.* (1996) between log cumulative practice alone (deliberate practice) and current Elo rating (chess-skill rating measure) for chess players under 40 year of age. A log scale was used to represent the effects of case reading experience because several investigators have suggested that the relationship between practice and learning is best expressed by a power function and this has been referred to as the Power Law of learning (e.g., Newell and Rosenbloom, 1981; Anderson, 1995).

The range of case-reading experience in Figure 19.1 was from about 1 log case readings to 4.1 log case readings, or about 10 to 12,000 cases. This range includes two residents at the beginning of mammography training with very little case reading experience (<1 log case) who performed at an A1 of about 0.500, where 0.000 is chance performance under the AFROC curve, and 3 mammographers with from 10,000 to 12,000 case reading experience (>4.1 log cases) who performed at A1 = 0.840. The training level of the observers is indicated by numbers or letters associated with the data points. Overall performance increases directly with experience, and in an orderly progression with training level. The fact that the beginning residents' performance is above chance at the start of mammography rotation can probably be attributed to reading experience from other specialities encountered during residency rotations as well as book reading and didactic sessions on mammography. Talent is also a factor that plays a role in the relationship shown in Figure 19.1, and shows the greatest variability in the third- and fourth-year residents who are nearing the end of their training experience.

The main point of Figure 19.1 is that logarithmic increases in mentor-guided mammography reading experiences are required to produce skilled mammography reading performance, and even at the end of mammography training, residents' interpretive skills are significantly below that of their mentors and will, according

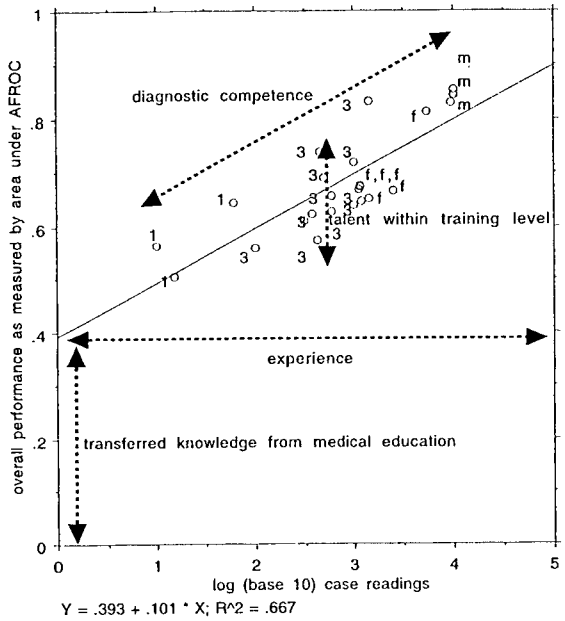


Figure 19.1: A regression analysis of overall performance measured as A1 as a function of log (base 10) number of cases read over a 3-year period by 3 experienced mammographers and 19 radiology residents undergoing clinical mammography rotation. When case readings is zero, the regression line intercepts the y-axis at 0.393 A1. With mentor-guided case-reading training and experience, A1 performance increases. The numbers next to the data points indicate the level of training and experience of the observers: 1 = first- and second-year residents; 3 = third- and fourth-year residents; f = fellows; and, m = mammographers. As indicated by the diagram below the data, competence increases with experience. Differences within levels are assumed to be due either to talent or random variation.

to this plot, require massive further amounts of reading experience. The interesting question is whether other forms of training more closely aligned to the notion of deliberate practice, which might be achievable by providing computer-assisted feedback as part of the training, would produce more effective learning. This makes the computer the mentor, but as such the computer can only be programmed to provide "plausible" feedback to student inquiries since image "truth" is unknown.

19.5 What is learned from reading medical images?

The usual answer to what is learned is knowledge, which is translated into various forms of cognitive skills and decision strategies. The knowledge skills that are

learned provide a perceptual basis for recognizing disease states in medical images and a cognitive basis for translating image perceptions into diagnostic disease categories. The details of these cognitive skills and strategies are elusive because the experimental means of getting at them is indirect and usually couched in cognitive theory. This is why expert systems have not generally led to practical results. If these skills and strategies cannot be identified, they cannot be taught. The simplest solution is to override the perceptual-cognitive analyses that attempt to identify the skills and strategies and instead resort to performing massed practice on the task that best represents the domain of expertise. It is agreed by many researchers from both perceptual and cognitive camps that massed practice is the main change agent in achieving expertise. Thus, if the goal is to be a radiologist, then the prescription for gaining expertise is to learn about radiology by practicing reading radiographs. Practicing reading radiographs has to be supplemented by feedback about whether the readings match reality in terms of diagnostic truth, and making appropriate adjustments (error-correction feedback). This means that training in medicine, particularly anatomy and pathology, as well as in radiology, which depends on 3D spatial abilities, is a necessary prerequisite. So the sequence for gaining expertise in radiology is: TRAIN & READ RADIOGRAPHS—SEEK FEEDBACK—ADJUST READING TO FIT DIAGNOSTIC FACTS.

We will use examples from the research literature on radiology expertise to compare the perceptual and cognitive approaches designed to answer the question, what is learned? The general framework for problem solving in the radiology domain for both approaches can be summarized as:
SEARCH & DETECT—RECOGNIZE—DECIDE.

In other words, three different aspects benefit from learning: developing a heuristic search and detection strategy, fine-tuning visual recognition of targets through practice, and balancing the likelihood of being correct against the possibility, and cost, of error in decision making. A recent version of the perceptual model is shown in Figure 19.2.

This figure shows that a global percept can be extracted from the image. This corresponds to obtaining an overall impression of what is being displayed. From this global percept, objects are separated and representations of disease-free areas are segregated from representations of possible-lesion areas. The lesion candidates are then scanned for feature extraction, which is the initial step in hypothesis formation by the observer. These features will work as guides to the expert, by suggesting the possible diagnostic outcomes. Once this diagnostic list is generated, it is confirmed against the features observed, which gives the expert a probabilistic distribution of the possible diseases. The possible disease list is checked with the objects perceived in the image, and a new search activated. In this way expert reasoning works in two directions, bottom-up, by carrying out object segmentation and feature extraction, and top-down, by checking the image elements against the diagnostic list.

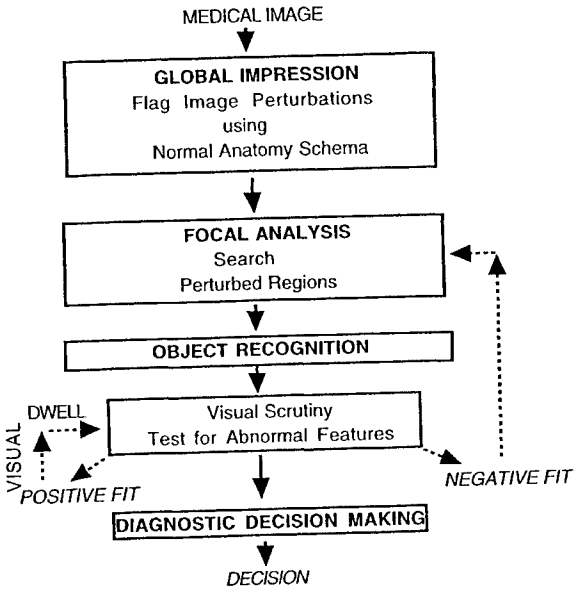


Figure 19.2: A perceptual model of the radiology task. The model shows the information-processing flow from the presentation of the image to diagnostic decision. Initially, a medical image elicits a global impression that flags perturbations setting up focal analysis in which perturbed regions of the image are searched. This results in recognition of objects that are tested for abnormality. The outcome of each test is either a positive or negative fit to the abnormality being tested. In either case, the testing is recursive. If a positive fit is found, the object is scrutinized by multiple eye fixations resulting in a build up of visual dwell in the region of interest. If a negative fit is found, attention shifts back to the medical image for a new global impression flagging another perturbed region, focal analysis searches it, a new object may be recognized and recursive testing for abnormalities continues until the observer is satisfied that enough evidence has accumulatied to make a diagnostic decision. Under this model, true abnormalities may be detected and receive fixation dwell but fail to be reported.

19.5.1 Search

A key question that drives both the perceptual and cognitive approaches is: How is search guided by knowledge (Newell and Simon, 1972), whether searching the visual display or searching the problem space for a diagnostic solution? The perceptual approach attempts to derive answers by analysis of eye-position data that search and test image features for diagnostic information leading to a decision outcome. The cognitive approach attempts to derive answers from analysis of verbal

protocol data that reveal cognitive structures embedded in propositional-statement logic referring to image findings used to generate diagnostic solutions.

19.5.1.1 Eye movements and searching the visual-image space

Searching with the Eyes—One of the earliest attempts to study the role of search in radiology expertise was carried out by Kundel and La Follette (1972). They were interested in the evolution of expert search patterns. Kundel and Wright (1969) had already provided evidence that radiologists frequently use a circumferential search pattern when searching for lung nodules in chest radiographs. The circumferential pattern reflects a heuristic search strategy for selectively sampling information on the radiograph based on prior knowledge about the type of target abnormality (e.g., lung nodule), or expectations about disease type (e.g., clearly recognizable multiple abnormalities). The evolution of a heuristic search was clearly demonstrated in a follow-up study which compared eye-fixation patterns of untrained laymen, medical students, radiology residents, and staff radiologists viewing normal and abnormal chest radiographs without prior knowledge about type of target abnormality. According to the authors, this heuristic search strategy evolved mainly as a result of "... knowledge of radiographic anatomy, pathology, and clinical medicine rather than upon formal radiologic training as given in residency programs." A specific form of the knowledge that guides search is "... clear and unambiguous definitions of 'normal' and 'abnormal' " (Kundel and La Follette, p. 528). This knowledge comes from years of experience reading chest radiographs to gain familiarity with features that distinguish targets of search from their anatomic backgrounds.

These early studies represent the beginnings of the perceptual approach to the study of expertise. They are important because they point out that radiology expertise is characterized by heuristic, not random, search. The term heuristic is popular in AI research, and books have been written arguing about its meaning and significance in AI (e.g., Groner, Groner, Bischof, 1983). We refer to heuristic search here meaning that experts choose an approach in searching a radiograph which draws on prior knowledge and experience to form an initial hypothesis that guides search rather than searching in a trial and error manner without preconceptual guidance. This strategizing is an interesting trade off that human observers choose in solving problems, in contrast to machines that typically use an exhaustive sampling of the problem space until the target of search is detected. A good example of this can be found in world champion-level chess programs which are capable of a 14-ply full-width search, where ply refers to one move and countermove. Contrast this brute-force search with skilled human chess players who typically look only one or two plies deep, even though they could look 8–10 plies deep (Charness, 1981). Humans are unwilling to expend the energy required to carry out an exhaustive search for the small amount of gain that it yields. Applied to searching radiographs, this translates into a search strategy in which the observer attempts to use the smallest effective visual field to sample the most informative image areas in a minimum

amount of time (Kundel, Nodine, Thickman, *et al.*, 1987). An expert uses structured knowledge to adjust the visual field size, determine the informative areas of the image, and keep track of the information yield over the time course of search.

Evidence that structured knowledge guides the search of experts comes from comparing random versus systematic-scanning (exhaustive) models based on human eye-fixation parameters. This comparison shows that radiologists confine their scanning to the lungs, in a chest radiograph, with a visual field between 2 and 3 deg (radius), searching for lung nodules. We speculate that gobal-image properties define the boundaries of the target-containing area. By 10 sec the radiologists have covered 85 percent of the lungs and detected most of the lung nodules. In the same time, the exhaustive model has covered more of the chest area, but not more lung area. The search pattern of the radiologists has been shown to exhibit more consistency with a circumferential pattern most common when searching for lung nodules, but for more general search tasks consistency gives way to idiosyncratic patterns that are too complex to categorize (Kundel and Wright, 1969). These findings suggest that scanning patterns of radiologists are not random but rather dependent on what the radiologist perceives to be the task, and what is seen during the course of scanning the image.

19.5.1.2 Verbal protocols, thinking out loud and searching the cognitive-problem space

Searching with the Mind—The cognitive approach has downplayed the visual component of radiology expertise and focused on the observer's cognitive evaluation of the radiographic display. The approach is similar to that used by Chase and Simon for studying chess in that a perceptual phase and a cognitive phase are separately tested. As an example, we use the experimental protocol of Lesgold, Feltovitch, Glaser *et al.* (1981, 1988). First the observer is given a brief view of the radiograph (2 sec), and asked what was seen with experimenter prompts to test the limits of the initial perceptual phase. Second, a verbal protocol is elicited by having the observer read the radiograph while "thinking out loud." This is followed by a formal dictated diagnostic report. Finally, the observer is given the patient history, re-examines the radiograph, and if necessary makes modifications in the diagnostic report. The goal of the perceptual phase is to identify how the stimulus is initially represented and schematically encoded within the problem space and to get at tentative hypotheses. The goal of the verbal protocol phase is to identify reasoning paths to a diagnostic solution. The reasoning step was modified in the experimental protocol by dropping the initial perceptual phase and expanding the verbal protocol phase to include having observers circle key areas on the radiograph that were considered critical in the reasoning path to diagnosis.

Lesgold, Feltovitch, Glaser *et al.* studied expertise in chest radiology by comparing verbal protocols of radiology residents at various levels with those of experienced radiologists. Analyses of verbal protocols led to a two-stage model of the diagnostic process. In the first stage a perceptual decision is generated. This yields a probabilistically-weighted set of perceptual features that diagnostically

characterizes the radiographic image leading to a single outcome. The initial stage is followed by a decision-making analysis of the perceptual features within a cognitive framework of diagnostic problem solving (see Selfridge, 1959). The first stage depends heavily on the observer's schematic knowledge in the form of an anatomic representation—a map of chest features. The second stage depends on cognitive testing of perceptual features that are translated into radiologic findings used to feed diagnostic reasoning chains. Radiology expertise was reflected by a richly structured anatomic schema mapping x-ray features to normal anatomy. This rich schema provides the basis for detecting "left over" features signaling and localizing possible abnormalities on new chest x-ray image instances. A schema is called up faster in experts than trainees giving experts a faster start in searching the radiograph and a more accurate roadmap of abnormal features likely to trigger a decision-making rule leading to a diagnostic solution.

The skill component of expertise was demonstrated by the fact that experts exceeded trainees on all quantitative measures derived from protocol analysis (e.g., more findings, more and bigger clusters of findings, more relationships among findings, and more inferential thinking using findings). Experts were also better than trainees at recognizing and localizing perturbations in normal anatomic structures that signaled pathology. Analysis of protocol statements emphasized the problem-solving flexibility of experts in generating schemata to fit specific cases, holding them tentative and accepting or rejecting them only after rigorous testing. The experts seemed to be able to generalize the x-ray findings from specific cases to idealized patterns of disease by drawing on mental models of patients' anatomy and medical history. Trainees showed less flexibility, generating schema so tightly bound to perceived x-ray findings that they often led to false solutions. For Lesgold, Feltovitch, Glaser *et al.*, the schema is the key to successful problem solving and "... acquisition of expertise consists in ever more refined versions of schemata developing through a cognitively deep form of generalization and discrimination." (p. 340). For most radiographic diagnoses, a shallow level of cognitive processing will suffice, and may even be more accurate than deeper reasoning (Proctor and Dutta, 1995). It is only with complex diagnoses that the advantage of deep cognitive processing becomes apparent. This deep form of cognitive processing comes about as a result of extended practice that makes expertise in radiology possible. Raufaste, Eyrolle, Marine (1998) expand on this conclusion by testing what they call a "pertinence generation" model of radiology diagnosis. Protocol analysis of 22 radiologists' interpretations of two "very difficult" cases revealed two qualitatively different kinds of expertise, basic and super. The super experts were distinguished by increased pertinence in the interpretation of diagnostic findings. Pertinence generation refers to linking visual signs (radiologic findings) to diagnostic inferences which increases with level of expertise. The cognitive processing behind pertinence generation is schema driven but the reasoning chain is more deliberate and reflective in super experts than basic experts or radiology residents. Super expertise is not simply acquired through more and more experience. Rather, at least for Raufaste, Eyrolle, Marine (1998), super expertise is the integration of reading experience with teaching and research experiences. These provide

a basis for integrating understanding about how radiologic findings are translated into plausible (pertinent) diagnostic hypotheses and logically tested for pathologic process leading to a diagnosis in much the same deliberate manner as deductive reasoning is carried out in a scientific experiment.

19.5.2 Visual recognition—features vs patterns vs objects

We have talked a lot about the importance of schemata in radiologic problem solving. Neisser (1976) built his theory of cognition around the concept of anticipatory schemata. According to his view, perception is a constructive process of discovering what the visual world (or image) is like and adapting to it. This discovery and adaptation process is, in the most general sense, the goal of visual information processing. It is possible to view radiographic diagnosis as a constructive process. For Neisser, the perceptual cycle is elicited by a schema that directs visual exploration to sample objects (information) and feed back the results thus modifying and enriching the initial schema. A schema for Neisser defines plans for perceptual action and readiness to take in certain kinds of perceptual structure. If an evoked schema is to be effective for guiding search, it must be generated early in problem solving. How does the initial visual input from the radiograph stimulate the formation of a schema? After initial schema formation, what role does visual recognition play in evaluating targets detected during focal search? What is recognized, features, patterns, objects or what? Both perceptual and cognitive approaches have focused on features as the basic unit of cognitive processing. This leads to a bottom-up analysis by synthesis of the object to be recognized. Both approaches also talk about the importance of patterns and chunking of information which are higher-level perceptual or cognitive structures. David Marr (1982) pushed visual recognition to the top-down object-representation level, and this has been expanded by Ullman (1996). The box that one gets into in postulating models of the visual recognition process is the chicken-and-egg dilemma: whether the observer first detects a distinctive part (feature) and builds up the object percept; or, whether the object percept is globally recognized, holistically, without intervening building-block steps. This is a critical question for cognitive modeling underlying visual recognition because we build error-correcting feedback based on our theory of the information-unit building blocks. Thus, if our theoretical building blocks are features, we train by feeding back features of object to-be-recognized. And how do we confirm that the features are truly the building blocks behind visual recognition? We ask the observers to think out loud and they say they use FEATURES to recognize the object. This circular logic is prevalent in the theoretical accounts of both perceptual and cognitive approaches used to study visual recognition that are reviewed below, and the answers they generate. Different experimental methodologies have been used to try get at the answers to these questions, but visual recognition still remains a puzzle.

Flash experiments. One answer to the visual-recognition puzzle comes from so-called flash studies in which radiographic images are presented briefly (e.g., 200 ms, the typical duration of an eye fixation) using a tachistoscope (Kundel and

Nodine. 1975; Gale, Vernon, Miller *et al.*. 1990). Several studies have asked how much can be seen in a single glance by tachistiscoptically presenting radiographs to radiologists and asking for diagnosis. In 1975, Kundel and Nodine identified a key skill that characterizes radiology expertise. They found that in 200 ms experienced radiologists could accurately recognize 70 percent of the abnormalities detected under free search of a chest radiograph. Many of the abnormalities recognized in 200 ms were large, high-contrast targets (e.g., mass, pneumonia and enlarged hearts) which significantly altered the appearance of normal anatomic structures in the chest x-ray image. Small or low-contrast targets (e.g., lung nodule, histoplasmosis) were not detected in 200 ms. This led to the interpretation that a global response (akin to Gregory's "object hypothesis." 1970) involving input from the entire retina provides an overall (schematic) impression of the radiograph that initiates focal search to test image perturbations leading to a diagnostic decision (Kundel and Nodine, 1983). The initial grasp of the visual scene is compared against schemata in which stored knowledge representations and deviations from expectations of a normal chest pattern are spatially encoded and flagged, all within the average duration of a single eye fixation. Kundel and Nodine (1983) showed that differences between radiologists' and laypersons' schemas of radiographic images are reflected in their drawings of what they saw. The drawings by laypeople, who did not recognize what they were looking at, consistently depicted background features. The drawings by radiologists, who did recognize what they were looking at, depicted image objects (see Figure 19.3). Interestingly, when looking at a hidden figure (puzzle-picture of the head of a cow), unfamiliar to both groups, the drawings leading up to recognition consistently focused on background features surrounding the hidden object for both groups. Only after the cow was recognized, presumably as the result of a match to an appropriate cognitive schema, did the drawings depict the hidden object (see Figure 19.4). Correlated with shifts in focus of the drawings from features to objects was a corresponding shift in focus of eye fixations from background features to object centers.

Similar findings have been found for detecting breast lesions (Mugglestone, Gale, Cowley *et al.*, 1995). For example, Mugglestone, Gale, Cowley *et al.* compared mean overall performance (Az area, that is the area under the Receiver Operating Characteristic or ROC curve), mean percent correct recall and mean percent return to screen of 9 radiologists under flash (200 ms) and unlimited viewing. They found overall that performance was poorer under flash than unlimited viewing (0.518 vs 0.700, respectively), primarily attributable to missing subtle abnormalities that failed to standout against mixed and dense breast-parenchymal backgrounds compared to lesions in fatty breast backgrounds (49% vs 31% for flash and unlimited viewing respectively). The lower overall performance for breast images compared to chest images was primarily due to the interaction of the conspicuity of the abnormality with anatomic background structures in the image. Inconspicuous, solitary findings in both breast and chest images seem to lack perceptual saliency in flash viewing. It may also be that lack of anatomic landmarks in breast images may fail to provide a distinctive anatomic schemata to facilitate pattern recognition compared to chest images which are rich in anatomic structures. Reflections

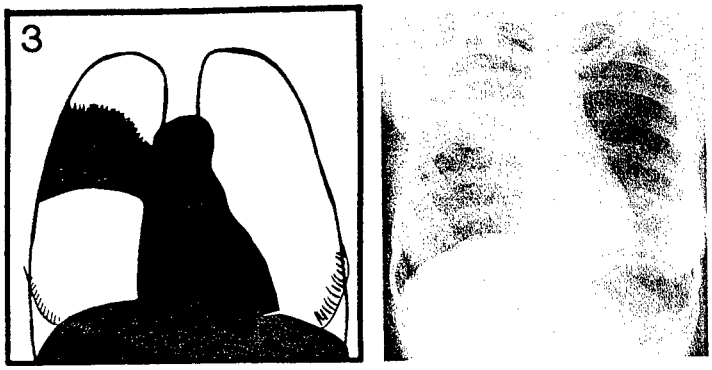
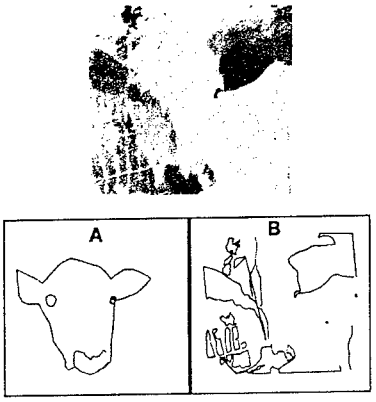


Figure 19.3: A silhouette drawing of a chest x-ray image containing a right-upper-lobe pneumonia (left), and the actual chest x-ray image containing the abnormality (right). The silhoeutte drawing was made by an experienced radiologist to illustrate the major finding on the chest x-ray film which was used in a flash study of image perception (Kundel and Nodine, 1975). The drawing depicts the abnormality in schematic fashion which may reflect how an experienced radiologist's cognitive schema encodes the chest-x-ray image when viewed in a 200 ms. flash presentation. The actual chest x-ray image was used in the experiment. In flash viewing, 30 percent of the observers gave an accurate diagnosis. In free viewing, diagnostic accuracy increased to 50 percent.

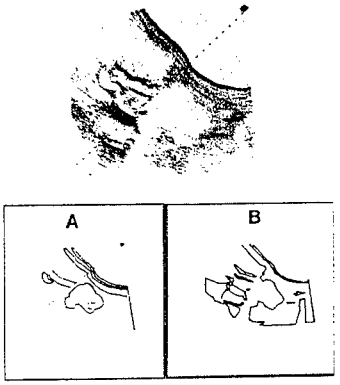
of these schemata are clearly illustrated in the drawings of chest-disease patterns shown in silhouette form in Kundel and Nodine (1975).

This would mean that the initial global impression for a breast image would key on conspicious features rather than anatomic landmarks, and that the global impression would be less effective in guiding focal search of the breast image. It is doubtful that the global impression can detect microcalcifications, and this was confirmed by Mugglestone, Gale, Cowley *et al.* (1995). Maybe this is why we observe that mammographers typically make two passes over a case that they are reading, the first to gather an overall impression and check for masses, and a second slow deliberate scan with a magnifying glass to catch microcalcifications.

Decision-time experiment. A second answer to the visual-recognition puzzle comes from a decision-time study of expertise in mammography which shows that experts are significantly faster and more accurate in detecting breast lesions than less-expert observers (Nodine, Kundel, Mello-Thoms, *et al.*, 1999). The initial detection, localization and classification of true lesions by experts occurred within 15 sec on average. This is much longer than flash viewing but in this case decision time included search time scanning both craniocaudal (CC) and mediolateral oblique (MLO) breast images for lesions, and detection plus localization time using a mouse-controlled cursor. The speed and accuracy of expert performance suggests to us a rapid global image impression that cues efficient focal search and supe-



(a)



(b)

Figure 19.4: A puzzle-picture of the head of a cow (top, Figure 19.4(a)), and a longitudinal ultrasound image of the abdomen showing a dilated bile duct and the head of the pancreas (top, Figure 19.4(b)). These images were shown to 6 observers, 3 radiologists and 3 laypeople. The outline drawings under the pictures show what observers saw after 20 sec viewing. Observer A in Figure 19.4(a) (lower left) reported that he saw a "cow's head." Observer B in Figure 19.4(a) (lower right) reported that he saw an abstract picture of a "fish." The outline drawings below the picture in Figure 19.4(b) show what the observers saw after seeing the ultrasound image. Observer A in Figure 19.4(b) (lower left) was a radiologists, and reported seeing a "dilated common duct" from a pancreatic mass. Observer B in Figure 19.4(b) (lower right) was a layperson, and reported seeing an "aerial photograph." The drawings suggest that visual concepts of what observers thought they saw are driving image perception.

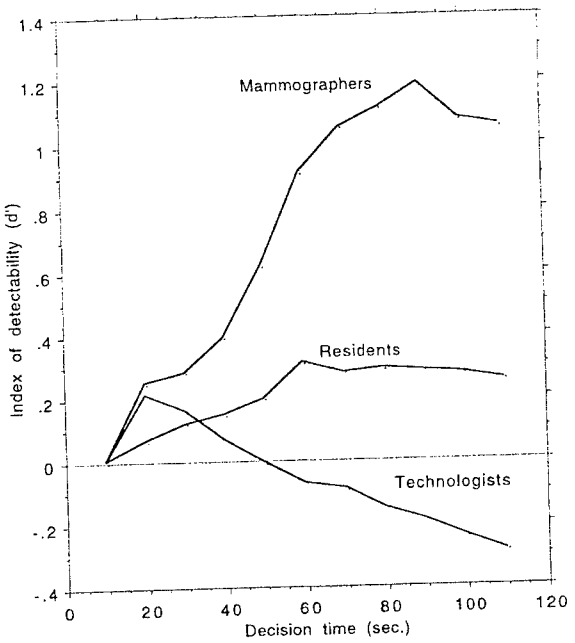


Figure 19.5: Speed-accuracy relationship as indicated by d', the index of detectability, as a function of decision time for mammographers, residents and technologists performing a combination mammography screening-diagnostic task (Nodine, Kundel, Mello-Thoms *et al.*, 1999). Overall performance as measured by d' which is the normal deviate, z(TP), of true positive fraction – z(FP), of the false positive fraction, increased for mammographers and to a lesser extent for residents. Overall performance decreased below chance ($d' = 0$) for technologists, meaning that false positives outnumbered true positives. Differences in performance were hypothesized as primarily due to lack of perceptual learning, which limited object recognition skills, causing competition between true malignant lesions, benign lesions, and normal image perturbations.

rior visual recognition of lesions. Initial impression, search and evaluation were more drawn out in observers with less expertise, and breakdowns in performance resulted in fewer true positives and more false positives. Figure 19.5 shows the speed-accuracy relationship related to mammography expertise.

Eye-position experiments. A third answer to the visual-recognition puzzle comes from eye-position studies of expertise in mammography (Figure 19.6) which also show that experts are faster and more accurate at detecting breast lesions (Nodine, Kundel, Lauver *et al.*, 1996; Krupinski, 1996). Using time to hit (TTH, search time to fixate a lesion) as the dependent variable, these studies show that experts

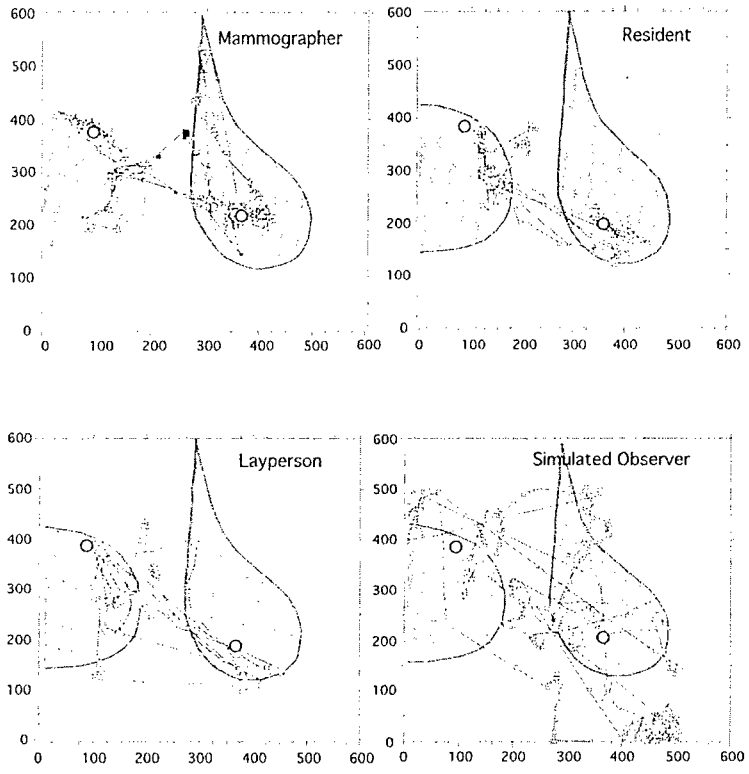


Figure 19.6: Scanning patterns of a mammographer (upper left), resident (upper right), layperson (lower left), and the simulated observer (lower right) to a digital mammographic image containing craniocaudal (CC) and mediolateral oblique (MLO) projections. The duration of the scan was limited to 16 sec in this comparison. The center of the mass in the CC view (left), and MLO view (right) are indicated by circles. The mammographer hits the mass in both views within 2 sec. The resident fixates the mass in the CC view in 1 sec, but takes almost 16 sec to fixate the mass in the MLO view. The layperson comes close to fixating the mass in both views. The simulated observer performs a random walk for 16 sec ultimately fixating the mass in the MLO view, but missing the mass in the CC view. The simulated observer scanpath uses saccade length and gaze duration data from human scanpaths. This results in fixations clustering in several regions of the display, but these clusters are independent of image content, since fixation x, y locations were randomly generated. Further details can be found in Nodine, Kundel, Lauver *et al.* (1996).

(experienced mammographers) find lesions faster than observers with less experience and training. Nodine, Kundel, Lauver *et al.* found that mammographers searching a two-view mammographic display containing CC and MLO images first fixated a mass that was reported correctly with an average TTH = 2.69 sec, whereas radiology residents required an average TTH = 4.74 sec to detect a correctly reported mass. Support for the view that the search strategy of experts was not random comes from a comparison of their TTH data with that of simulated observers that searched the breast image randomly.

Interestingly, the simulated observers took an average TTH = 4.67 sec which meant random fixations first hit an area containing a true mass after about the same search time as radiology residents. However, human observers failed to fixate only 2 percent of areas containing true masses whereas simulated observers failed to fixate 44 percent of areas containing true masses. These comparison data provide strong evidence that speed and accuracy of expert performance is tied to the rapid generation of a diagnostically-useful initial schematic representation that is effective in guiding search. We speculate that what experts recognize at first glance are unexpected oddities generated from a global characterization of the image that are flagged as regions to-be-searched by focal scanning. Thus, the goal of the initial global problem representation in radiology is not to find a target per se (because there are too many possibilities), but rather to find something odd about the image on which to focus the search strategy. Expertise comes into play in characterizing what in the image is odd. To recognize this the observer must first know what is not odd, or what is "normal." Evidence that the visual recognition of experts is tuned to differentiate odd or uncharacteristic features signalling pathology from clinically normal features comes from Myles-Worsley, Johnston, Simons *et al.* (1988). These odd features that occur in x-ray images have been called "perturbations" implying that their presence in the image disturbs the observer's image representation (schema). They found that as radiologists develop expertise in recognizing clinically-relevant abnormalities, they tend to selectively ignore normal feature variants, suggesting that detection of perturbations becomes more refined. Both perceptual and cognitive approaches agree that one of the most important signs of expertise is speed and accuracy of recognizing globally whether an image is normal or abnormal. This preceeds detailed search and analysis which leads to a specific diagnosis, and even this phase is faster in experts.

Probability-analysis experiments. A final answer to the visual-recognition puzzle comes from probability analysis of error paths in breast lesion detection (Mello-Thoms, Nodine, Kundel, 1999). What this analysis shows is that the initial decision made when examining a pair of breast images (CC and MLO views) significantly influences any subsequent analysis on that image. Namely, when the first decision is a true positive then the probability that the observer will find the same lesion on the other view is very high, with experts being significantly better than residents or technologists. Furthermore, on average, in this senario, the observer will make significantly fewer mistakes (false positives or false negatives) than if the first decision is incorrect (false positive). Moreover, when the first decision is

incorrect, then the probability that the observer will find the true lesion, when one is present, is very small. Thus, beginning with error seems to promote other errors, dragging performance down. Interestingly enough, when the first decision is a false positive, in an image pair that has a true malignant lesion visible, on average the observers will make significantly fewer errors than when the first decision is a false positive in a normal image pair. In fact this difference can be quite staggering depending on the level of expertise. Experts will make about 8 times more errors when the first decision is a false positive on a normal image than when it is a false positive on an image with a malignant lesion present, whereas residents will make about 5 times more errors and technologists will make only about 2 times more errors. Maybe this is because of the confidence that experts have in their decisions, or because the image perturbation that led the expert to make the initial false positive decision on a normal image repeats itself in other areas of the image, thus misleading the expert into making other incorrect decisions. With residents and technologists this occurs to a smaller degree, probably because these two groups generate more errors on a regular basis, that is, they are more consistently fooled by image disturbances. In other words, the presence of a true lesion, even when the true lesion is not reported by the observer, seems to work as a perceptual bias for the number of false positives made.

Both the perceptual approach and the cognitive approach stress the importance of a rapid initial mental representation of the problem. Whether this is referred to as a schema or cognitive structure makes no difference because both perceptual and cognitive approaches are referring to representations of the same process. The perceptual approach uses visual-feature mappings, and the cognitive approach uses logical rule-based mappings to represent problems and generate solutions. The flash studies show that experienced radiologists have clear and unambiguous definitions of "normal," from which fast accurate recognition of deviations are globally detected.

19.5.3 Decision making

Evidence for differences in decision making as a function of level of expertise comes from two sources. First, from eye-position studies of observers who make perceptual errors in radiology, and second from our study of the speed-accuracy relationship in developing mammography expertise (Nodine, Kundel, Mello-Thoms *et al.*, 1999).

Eye-position studies have identified three kinds of error in lung nodule detection: search errors, detection errors, and interpretation errors. Two-thirds of errors are divided between detection and interpretation, not search (Kundel, Nodine, Carmody, 1978). Visual dwell data show that missed targets (breast or chest lesions) receive as much if not more visual attention as do recognized, truly-positive targets (Krupinski, Nodine, Kundel 1998). This means that observers look at the missed target long enough to report it, but decide not to report it. Thus, over 60 percent of missed targets seem to be cognitively processed, as evidenced by both fixation clustering and prolonged visual dwelling on the missed target, yet observers fail

to find sufficient evidence to report the object they evaluated as a target candidate. Analysis of eye fixations and visual dwell provides an information-theoretic account of the cause of errors of omission, false negatives, in radiology. Unfortunately it is difficult to disentangle whether the cause of the omission error was faulty recognition or decision making. Errors of commission, false positives, are also associated with prolonged dwell times that are equivalent to those found for true positives.

To shed light on this we looked at overall performance (area under AFROC) as a function of the time course of viewing mammographs by observers representing different levels of mammography expertise (Nodine, Kundel, Mello-Thoms *et al.*, 1999). We measured decision time, which is equivalent to what experimental psychologists refer to as "reaction time" (Posner, 1986), and related it to decision outcome using a combination mammography screening/diagnostic task. We have already reported above that experts were faster and more accurate performers, and that this is attributed to a well-developed prototypic normal breast schema that facilited the recognition of abnormal deviants correctly evaluated as malignant lesions. Perhaps the most interesting finding coming from this experiment is not the speed-accuracy relationship of experts, but rather that of the least expert in mammography interpretation, the mammography technologists, who had neither training nor experience reading mammograms. One technologist stands out in particular because she took the task literally and called every visible blob on every case. Her decision criterion for deciding that a malignant breast lesion was present was: Do I see a blob? She called 193 malignant lesions on 150 breast images of which 50 (26%) were correct. Her strategy appeared not to be driven by a schematic representation that maps anatomic knowledge with pathological knowledge. Rather, her strategy was driven by a simple blob-detection algorithm. In comparison, an expert (mammographer) called 97 malignant lesions of which 52 (54%) were correct. Thus, the mark of expertise is not how many correct lesions are recognized, but rather the balance between reporting true lesions and minimizing reporting false lesions. This calls on highly-tuned perceptual discrimination and differentiation which is learned through massive amounts of image-reading experience supplemented by feedback.

19.6 Connectionism—another approach to information processing

The perceptual approach and the cognitive approach to information processing deal with the reasoning process leading to decision making in very different ways. In particular the cognitive approach attempts to create a set of rules that will guide perception, evaluation and decision making. A shortcoming of this method is that different experts in the same field may have very different reasoning processes, as shown in Lesgold, Feltovitch, Glaser *et al.* (1988), Raufaste, Eyrolle, Marine (1998). This imposes tremendous difficulties to modeling the decision process, because input from each expert has to be carefully weighted and placed in the reasoning steps of a model that a computer can execute.

The cognitive approach can be seen in the first generation of AI systems, based upon predicate logic, which implied that all of the conclusions had to be drawn from a set of logical statements that basically corresponded to a game-playing scheme. Even the attribute-based representation systems, which allowed for broader mappings than the predicate logic did, used some sort of rule at each step to arrive at any conclusion. This was particularly true regarding decision trees. In predicate logic and attribute-based representations, sets of rules are designed based upon experts' verbal explanations of their analyses of medical images. This approach, however, impaired the use of artificial intelligence in problems for which no such set of rules could be consistently derived. Nonetheless, this does not imply that systems with this kind of representation are doomed to failure. There are obvious applications in which such a set of rules can be derived, and the problem is thus successfully solved.

An interesting point that can be made regarding these knowledge representations is, are human internal representations like these? That is, do humans use some logic- or attribute-based approach to solve their problems? Furthermore, these approaches seem to indicate that a serial structure is necessary, because each new conclusion can only be drawn based upon the answer to the previous question. This serial approach to brain function has been contested (Barrow, 1996). It has been shown that this approach maps the brain to a universal turing machine, which leads to restrictions regarding the speed of information processing, the robustness of the system and its lack of flexibility to deal with complex decision making tasks such as medical image interpretation (Dawson, 1998). Moreover, if indeed perception and cognition are based on a set of rules, shouldn't experts have a similarly structured set of rules? But that is not what one sees in practice, which may indicate that, although certain processes may be dealt with by the brain in this way, not necessarily all processes are analyzable in this fashion. Thus another type of knowledge representation must be considered.

The perceptual approach deals with the creation of an internal map that is based upon features that were visually extracted from the scene. This approach does not need to infer what this internal map looks like, for it only looks at the sequence of steps in and out of the internal map. It seems rational to ask an intelligent system to do this, namely, to build its own internal representation based upon a set of percepts extracted from the problem and then use these features to process the information by running through the internal map and producing a decision.

Thus, a third approach to information processing is created. Namely, it is based on the processing capabilities of the human brain, with its parallel weighted connections, that receives input and produces output, although the "how" is not entirely clear. This approach is called connectionism, and it is used by a fast growing branch of AI called artificial neural networks (or ANNs). In this chapter only one type of ANN will be considered, namely, the multi-layer perceptron (MLP), which is a multi-layer feedforward network (Haykin, 1994).

One of the major drawbacks with using the connectionist approach is that it is not clear which elements from the input patterns have a more significant contribution to the classification process. This is often called the credit-assignment

problem, because the "thinking" process of the network is done at an internal (and unobservable) level, in the hidden neurons, and no insights can be gained into what actually helped the network achieve the final result. Furthermore, even if the network derives a probabilistic distribution for the data, it cannot tell the user which distribution it is, which does not help the human understanding of the problem, although some methods have been recently developed to extract knowledge embedded in ANNs (Tickle, Andrews, Galea *et al.*, 1998).

Interestingly enough, this seems to be the way that the brain of a human expert works. For example, one may ask a distinguished radiologist how he or she arrived at a particular diagnosis, but often times they will not be able to list all of the steps that they took, or which factors weighted more heavily than others to generate a conclusion. Obviously if the problem is simple (for example, if a large malignant lesion covers a portion of the breast) then one has no doubts about what generated the particular diagnosis, but in these cases the ANNs also perform quite well, because the weight of the evidence in one dimension (in this case, size) is so overwhelming (Found and Muller, 1996). These are often not the cases in which one is interested. The secret for good performance, particularly in tasks like cancer detection, lies in the subtle lesions, in the early findings that may prevent a starting cancer from taking over.

Nonetheless, acquiring expertise in radiology requires massive amounts of practice, which is a problem for the novice radiologist or for a radiology trainee. As previously discussed, performance improves as a function of deliberate practice, that is, of self-motivated practice, as long as feedback is available to correct errors. Most of the time it is not possible to use a human expert to provide this feedback to inexperienced radiologists. In this sense the use of intelligent systems to aid these practitioners seem quite logical. The intelligent system can work as an educated second opinion, or by providing feedback to the observer about specific regions in the image.

In this section we will briefly discuss intelligent systems and artificial neural networks, as well as examine expertise in the context of ANNs. We will also discuss how to compare the performance of human experts with that of their artificially intelligent counterparts.

19.6.1 What is an intelligent system?

We will consider that an intelligent system is one that has agents (that is, elements) that allow it to successfully interact with its environment (Russell, 1996). Note that the definition as stated uses a measure of performance to determine if the system's actions in the environment lead to success or failure. It also assumes that knowledge about the environment is available to the system in a format that the system can not only use but that covers the universe of the domain of the problem. In other words, this knowledge is sufficient to allow the system to respond to its environment in an appropriate way (Partridge, 1996).

The interaction between an intelligent system and its environment as described above corresponds in psychology to a cognitive process (Fox, 1996), which is ap-

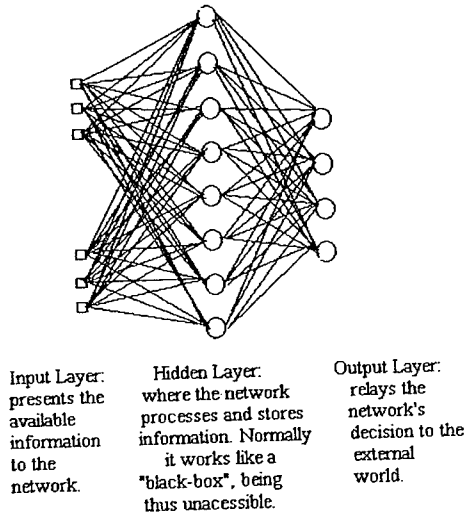


Figure 19.7: A generic representation of a multilayered perceptron. This network architecture is divided into 3 parts: the input layer, which receives information provided by the environment; the hidden layer (s), which processes that information; and the output, layer which transmits the network's decision to the environment.

propriate, considering that learning is one of the hallmarks of intelligence (Russell, 1996). In other words, if a machine can learn, then, in principle, it can become intelligent. In this context intelligence refers to the ability of freeing itself from its creator, namely, from making up its own hypotheses and assumptions about the environment, even if these contradict the original hypotheses that the system was taught (Russell, 1996). In the domain of radiology expertise, this means that the system should be able to find its own unique interpretation for a given image instead of trying to match it to the ones that were used to teach the system.

Only one type of ANN will be considered here, the multi-layer perceptron (MLP). This is a very powerful network architecture that has been proven successful in a variety of contexts (Haykin, 1994). A generic representation of the MLP can be seen in Figure 19.7. As shown, this network architecture is divided into 3 parts, namely, the input layer, which receives the information provided by the environment, the hidden layer(s), which processes this information, and the output layer, which relays the network's decision to the environment.

Multi-layer perceptrons can learn in two different ways. In supervised learning the system is presented with a set of examples and the truth table, that is, a list that maps each example to its correct category. Although this method has some obvious advantages—namely, feedback is immediate, because the system instantly knows

if it succeeded or if it failed—it nonetheless poses a problem in situations where no such truth table exists, that is, when there is no check on reality. In these cases unsupervised learning may be the best option. This learning technique allows the system to create its own map between "truth" and the examples presented to it. Thus, even though this is a more flexible type of learning, it has the drawback that many different classes may be created to represent objects belonging to the same category if the features that characterize these objects are not very similar. For example, many different classes could be created to represent masses in the breast, because these can be stellate, more or less dense, etc. Note that in this case, learning occurs by agreement, as opposed to by matching with a truth table as in the supervised learning case.

The process of presenting an ANN with a set of examples and letting it form its own representational map is called training. When previously unseen examples are presented to the network, its performance is judged by measuring how much and how well it learned. This process is called testing. If its responses are appropriate we deem that it learned to solve that particular problem. In this sense the learning process can be seen as a mapping between the examples domain, which offers discrete sampling about the (possibly) continuous multidimensional nature of the problem, and hypotheses formation, which allows the system to decide which action to take in the presence of a certain input. Note that in practice if the network performance during testing is below acceptance standards, it will have to be retrained. This process is equivalent to finding out that at the end of residency the performance of the residents is significantly below that of their mentors, and then attempting to improve their performance by exposing them to more cases.

Artificial neural networks have been successfully used in many areas of radiology, such as to predict breast cancer invasion (Lo, Baker, Kornguth *et al.*, 1997), to find calcifications in mammograms (Nishikawa, Jiang, Giger *et al.*, 1994), to detect signs of lung cancers in chest radiographs (Lo, Lin, Freeman *et al.*, 1998), and to differentiate benign from malignant lesions in mammograms (Zheng, Greenleaf, Gisvold, 1997).

Despite this success, it is important to consider that in most of the applications of ANNs to medical image reading, the selection of the features that will guide network diagnosis is done in one of two ways. Either image features and patient data are used to represent the problem to the network, or image parameters are extracted by some preprocessing step. Unfortunately, each of these representations has drawbacks. In the first case a human specialist has to search the image looking for the appropriate parameters (examples of such features could be the presence of calcification clusters, breast density, etc.), which may not be viable if the system is to be used to aid novice radiologists or to train residents, because neither of these groups may be completely capable of deriving such predictive features from the image. Furthermore, if one needs a specialist to derive the predictive image features, then one may as well use that specialist to read the image itself, and thus skip the ANN altogether.

In the second case, exhaustive search of the image is done in order to derive the features for the ANN. As shown elsewhere (Kundel, 1987) experts do not search

like this, but rather, use their prior knowledge (acquired by experience) to guide their search in a heuristic fashion, thus avoiding spending time looking in regions of the image where lesions are unlikely. One can argue that this procedure prevents total coverage of the image, and thus should increase the rate of false negatives, resulting from the presence of lesions in parts of the image where the expert failed to look. But this does not seem to be the case (Kundel, 1987). Experience seems to allow the experts to mentally generate a probabilistic map of the likelihood of lesions in different parts of the image. In taking information from this probabilistic map, the expert is in fact optimizing search using as constrains total time spent reading the image and the value of finding true lesions against the cost of missing a true lesion. Studies have shown (Nodine, Kundel, Mello-Thoms *et al.*, 1999) that experts are very fast and accurate in finding pairs of lesions given two mammographic views of the same breast, whereas residents and radiology technologists lag far behind. Furthermore, because of their lack of formal training, radiology technologists do not seem to build such probabilistic maps, but rather use what we called a "shot-gun strategy." They exhaustively examine the image and call everything that looks blob-like. As a consequence, the same criterion for lesion detection is used everywhere in the image, despite the local changes in anatomy (and thus in contrast appearance attributable to x-ray transmission), and the different likelihoods that a lesion will develop in different regions of the breast (Haagensen, 1986). This generates many incorrect decisions.

One of the biggest problems with artificial intelligence (AI) is its inability to deal with the incorporation of prior knowledge in the formation of new hypotheses. Note that this hinders the search process, because the machine cannot build the probabilistic map that experts do. This forces the search to be performed using a "shot-gun" strategy, which generates many false positives, and the results may resemble the exhaustive "blob-detector" radiology technologists described above, which is not acceptable for a system that aims to help radiologists.

19.6.2 Expertise in the context of artificial neural networks

Learning is an important part of improving performance in a decision making task such as mammography. Nodine, Kundel, Mello-Thoms *et al.* showed that human performance in reading mammograms improves as a function of individual talent and the number of mammogram cases read. Furthermore, we showed that experts have seen the largest number of cases, and also that their performance is significantly better than that of either novices or laypersons. In other words, we showed that human performance improves as a function of practice.

It is very difficult to use this measure to characterize expertise in artificial neural networks, primarily because in the vast majority of ANNs learning only occurs during the design part (a.k.a. training). This implies that, once the system has learned the input-output mapping with the examples provided to it, up to a desirable error level, it does not learn anymore. It is important to mention that this limitation is no fault of the theory of ANNs, but rather it is related to most of the algorithms currently available to train them. It is not impossible to develop a learning algorithm

that permits the network to learn continuously, and in fact such algorithms do exist (e.g., the adaptive resonance algorithm). The problem is that in order to allow the network to continuously learn one has to allow for the network structure to be flexible, that is, one has to permit the network architecture to change as new classes are learned. In mathematics this is called the "stability-plasticity dilemma," because a compromise exists between how much the network architecture can change and how these changes may affect the network's stability. This problem affects all of the existing algorithms to train ANNs, and it does not have a closed-form solution. Thus, the price that one pays for keeping the network learning is the risk of either making it so big that it takes a very large amount of time to generate an outcome, or having it become unstable. As a consequence of this, the most widely used algorithms do not allow the network to learn anything new, once it has been trained.

This impacts the network's performance in two different ways. Namely, the level of error generated when the network was tested in the laboratory or elsewhere represents the best level of error that the network is ever able to achieve, primarily because the testing samples were drawn from the same population that the network was trained on and under the same conditions (i.e., film quality, image acquisition setup, acquisition technique, digitization, etc.). Second, if the conditions are changed—for example, if an intelligent system is being used for cancer detection and the incidence of cancer in the population it addresses changes for some reason—then the majority of ANNs cannot adapt on their own to the new conditions, unless they are retrained taking into account the new situation. This is undesirable, considering that finding the appropriate set of parameters for an ANN may take anywhere from a few hours to a few months or even years.

One important point to consider at this step is that there is a minority of artificial systems that can respond to perceived changes in the environment on their own. An example of this would be the adaptive resonance theory neural network, which is an unsupervised learning network capable of creating new nodes to represent the new classes it encounters at any point in time, during or after training. Although this is a great advantage as far as adaptation goes, it is important to remember that it too suffers from the drawbacks of the unsupervised learning systems, namely, it may create many unnecessary classes in response to the variations in the input patterns.

A consequence of the fact that the ANNs can only learn during training is that it possesses a static knowledge, whereas the human experts possess a dynamic knowledge. All that the network knows today it will know tomorrow, but no more, although the human experts will continue to acquire knowledge. This greatly impacts the nature of expertise that the ANN possesses, namely, it is a different kind of expert than the humans, because its expertise is unchanging.

At this point, an important question remains: how does one measure the performance of an ANN? And, how does one contrast it with the performance of the human experts?

A methodology to measure the performance of an AI system versus the performance of human observers was proposed in Haynes (1997). In this case the AI

system would perform a task (reading a set of mammograms, for example) and humans with different experience levels would also perform the same task. A panel of expert judges would then rank performance, placing on the top of the list the better performances (more true lesions found, less mistakes, etc.) and on the bottom, the worse. The worst performance gets assigned a low score, and the best a high score. Observers with the same level of experience have their scores averaged, so that only one score represents each level of experience. A "skill function" is then plotted, which draws observer's experience versus the scores they received. Note that in this case the AI system is considered to be another observer. Confidence bands are also derived. Thus, to estimate the level of performance of the AI system versus the human experts, one can look at the plot and compare the AI's performance with that of the human observers that are closer to it. In this way one can make assessments such as "the ANN performed at a level of an observer with x years (or number of cases read, or whichever other measure) of experience."

One problem with such argument is that observers with the same level of experience may perform very differently, according to observer's talent, as shown in Nodine, Kundel, Mello-Thoms *et al.*, 1999. Thus, by averaging them together one is in fact misrepresenting performance at that level of experience. If, on the other hand, one uses only one observer with a given level of experience, then one is certainly risking representing that level with either the best or the worst performance, which certainly is not an acceptable measure.

Which criterion can be used then to measure the performance of an ANN? Well, certainly if one uses the same data set to test different ANNs, the one that has a smaller error is to be said to be the best. However, could one then go out and use this same data set to test human observers, and then compare these results with the ones from the ANN? The immediate answer is no, because human performance varies greatly with the level of expertise, which involves both talent and training. Thus, by saying that the ANN performed "better" than the human observers, one is in fact saying that it performed better against those particular observers which had a given level of talent and training. There are no guarantees, however, that as the number of cases seen by the observers increases, their performance would stay at the same level, and thus such a comparison is limited to that particular instant in time. On the other hand, if the human observers performed better than the ANN, nothing could be said, especially if the human observers had seen a larger number of cases than the ANN was trained with. One could say that the human observers in this case were trained with a larger (and possibly broader) data set. If the humans and the ANN had seen a similar number of cases, then one would have to be careful with the conclusions drawn, because one would have to show that this number is enough to generate human expertise (or to account for a decent training set for the machine).

What can be said, then, about expertise of an artificial neural network? One important point to consider initially is that, if all that constitutes expertise is something that can be computed, then there necessarily exists a set of rules that lead to it, because every mathematical identity can be rewritten as expressions from first-order logic (Bringsjord, 1997). Thus, if expertise is at all computable then ANNs

or any other artificially intelligent system can conceptually be capable of realizing the same type of expertise as humans, and probably of achieving as good or even better performance then human experts, because of their massive computational capabilities. If on the other hand, as we propose, expertise has a component that is not computable, such as the spontaneous generation of new concepts, then artificial systems cannot simulate that component, and their expertise will be different, in kind, to that of the human expert.

Another important part of human expertise is creativity. Namely, the ability to respond appropriately to the unknown, to derive a meaningful set of actions by contrasting the novelty with what is known, is one of the hallmarks of the human experts. In this way, creativity is another characteristic that separates the human expert from the novices. In this context all three different types of creativity (Boden, 1998) are to be considered, namely, the exploratory type, which involves the generation of new ideas by the exploration of the knowledge domain; the combinatorial type, which generates new ideas by creating new associations for old ideas; and the transformational type, which involves transforming what is known to generate a concept never before conceived. By possessing one (or more) of these types of creativity a human expert can not only further the knowledge in his/her domain of expertise but also derive a meaningful strategy once he/she is faced with an unknown (or never before seen) aspect of the problem, and then learn from it (Palmer, 1997).

Intelligent systems have a great deal of difficulty dealing with the concept of creativity. As pointed out elsewhere (Boden, 1996; 1998) only the exploratory type of creativity has been dealt with in AI systems, with a small degree of success. The other two types rely heavily in the human associative memory, and most of its processes are, as of yet, not completely understood and thus cannot be replicated in a network. As a consequence of this, the intelligent systems that currently exist cannot generate a new concept on their own, and when facing the unknown may not react appropriately, because of their incapacity to adapt to the new situation.

Thus, as a summary, we can say that we believe that intelligent systems should be more and more used to do things that people do poorly, because their capabilities for massive amounts of computations is very helpful in some situations. On the other hand, in tasks were people do well, the role of the expert system may be more restricted, such as that of a tutor or a peer whose second opinion should be taken into account, but that probably should not be left alone to run the show.

19.7 Conclusions

In this chapter we described the nature of expertise, particularly referring to expertise in radiology. We showed, for example, that mammography expertise, as measured by overall performance (area under the AFROC curve) is highly dependent on the logarithm of the number of cases read. Our recent study of expertise in mammography (Nodine, Kundel, Mello-Thoms *et al.*, 1999) also showed that residents in training develop similar decision-making strategies, as measured by

their use of decision confidence ratings, as expert mammographers. From a practical standpoint this suggests that resident training in mammography is effective in providing a general framework for learning radiology reading skills. But residents were inferior to experts in recognizing true breast lesions. We hypothesize that this weakness is primarily attributable to the lack of fine-tuned visual-recognition skills which are dependent on perceptual learning. Supporting the tuning of visual recognition argument, Sowden, Davies, Roling (1998) have recently shown that massed practice detecting calcifications in positive-contrast mammograms (bright target on dark background) positively transfers to a new task in which the calcifications are displayed in negative-contrast mammograms (dark target on bright background). This suggests that perceptual learning improves perceptual sensitivity in the detection of high-contrast targets. Massed practice was defined as a detection trial followed immediately by feedback about the correctness of observer's response. This improvement in perceptual sensitivity occurred even though the amount of massed practice was limited to 720 trials followed by a transfer test. The key to improvement seems to be the feedback. The development of expertise in chess playing, which draws on similar mental representations and optimization strategies to those for radiology expertise, also supports the importance of massed practice as the primarily change agent.

When we looked at the question of what is learned when reading medical images, we showed that acquired knowledge is translated into a variety of cognitive skills and strategies. As expertise is acquired, search strategies become less exhaustive and more probabilistically driven by enriched anatomic-pathologic schemas. Visual recognition of potential targets becomes more accurate because of an expansive image-reading repertoire that defines decision thresholds for normalcy. This acts to fine-tune discrimination and generalization thus facilitating perceptual differentiation of abnormalities.

We have proposed that expert systems possess a different kind of expertise than human experts, for they are only able to generate one of the two components of human expertise, namely, the computable part, called training. This by no means hinders their utility, but care should be taken when comparing the performance of an expert system to that of a human expert.

Most radiology expertise skills and strategies find representations in the three models of information processing discussed, the perceptual approach, the cognitive approach and the connectionist approach. Although each of these approaches deals with different representations of the same underlying system, they ultimately rely on the same basic learning pinciple about how the acquisition and processing of information occurs. The essential role of experience in learning is to enrich structured knowledge in order to facilitate radiographic interpretation

Finally, all the theorizing about how radiology expertise is acquired boils down to a very simple answer: Practice, which as we have defined it in this paper means case-reading experience, enriched by feedback in the form of knowledge of results, makes the structured knowledge perfect! This is not a very deep theory of learning, but it seems to capture the essence of how expertise in radiology is acquired.

Acknowledgment

The writing of this chapter was supported, in part, by Grant DAMD17-97-1-7103 between the USAMRMC and the first author.

References

Anderson JR. *Cognitive Psychology and Its Implications*, 4th Edition. New York: WH Freeman, 1995:304.

Barrow H. "Connectionism and Neural Networks." In Boden MA (Ed.) *Artificial Intelligence*. San Diego, CA: Academic Press, 1996:135–156.

Bass JC, Chiles C. "Visual Skill: Correlation with Detection of Solitary Pulmonary Nodules." Investigative Radiology 1990;25:994–998.

Boden MA. "Creativity." In Boden MA (Ed.) *Artificial Intelligence*. San Diego: Academic Press, 1996:267–292.

Boden MA. "Creativity and Artificial Intelligence." Artificial Intelligence 1998;103:347–356.

Bringsjord S. "An Argument for the Uncomputability of Infinitary Mathematical Expertise." In Feltovich P, Ford KM, Hoffman RR (Eds.) *Expertise in Context*. Cambridge, MA: AAAI Press/The MIT Press, 1997:475–497.

Charness N, Krampe R, Mayr U. "The Role of Practice and Coaching in Entrepreneurial Skill Domains: an International Comparison of Life-Span Chess Skill Acquisition." In Ericsson KA (Ed.) *The Road to Excellence*. Mahwah, NJ: Lawrence Erlbaum Associates, 1996:51–80.

Charness N. "Expertise in Chess: the Balance Between Knowledge and Search." In Ericsson KA, Smith J (Eds.) *Toward a General Theory of Expertise*. Cambridge: Cambridge University Press, 1991:39–63.

Chase WG, Simon HA. "Perception in Chess." Cognitive Psychology 1973;4:55–81.

Chi MTH, Glaser R, Farr MJ. *The Nature of Expertise*. Hillsdale, NJ: Lawrence Erlbaum, 1988.

Dawson MRW, *Understanding Cognitive Science*. Malden, MA: Blackwell Publishers Inc., 1998.

de Groot A. *Thought and Choice in Chess*. The Hague: Mouton, 1965.

Ericsson KA, Charness N. "Expert Performance." American Psychologist 1994;49:725–747.

Found A, Muller HJ. "Searching for Unknown Feature Targets on More than One Dimension: Investigating a 'Dimension-Weighting Account'." Perception and Psychophysics 1996;58:88–101.

Fox J. "Expert Systems and Theories of Knowledge." In Boden MA (Ed.) *Artificial Intelligence*. San Diego, CA: Academic Press, 1996:157–181.

Gale AG, Vernon J, Millar K, Worthington BS. "Reporting in a Flash." British Journal of Radiology 1990;63:S71.

Gobet F, Simon, HA. Templates in Chess Memory: A Mechanism for Recalling Several Boards. Cognitive Psychology 1996;31:1–40.

Gregory RL. *The Intelligent Eye*. New York: McGraw-Hill, 1970.

Groner R, Groner M, Bischof WF. *Methods of Heuristics*. Hillsdale, NJ: LEA, 1983.

Haagensen CD. *Diseases of the Breast*, 2nd Edition-Revised reprint. Philadelphia: Saunders, 1986:380–382.

Hayes CC. "A Study of Solution Quality in Human Expert and Knowledge-Based System Reasoning." In Feltovich P, Ford KM, Hoffman RR (Eds.) *Expertise in Context*. Cambridge, MA: AAAI Press/The MIT Press, 1997:339–362.

Haykin S. "Neural Networks: A Comprehensive Foundation." Englewood Cliffs, NJ: Macmillan College Publishing Company, Inc., 1994.

Krupinski EA, Nodine CF, Kundel HL. "Enhancing Recognition of Lesions in Radiographic Images Using Perceptual Feedback." Optical Engineering 1998;37:813–818.

Krupinski EA. "Visual Scanning Patterns of Radiologists Searching Mammograms." Acad Radiol 1996;3:137–144.

Kundel HL, Nodine CF, Krupinski EA. "Computer-Displayed Eye Position as a Visual Aid to Pulmonary Nodule Interpretation." Invest Radiol 1990;25:890–896.

Kundel HL, Nodine CF, Thickman D, Toto L. "Searching for Lung Nodules: A Comparison of Human Performance with Random and Systematic Scanning Models." Invest Radiol 1987;22:417–422.

Kundel HL, Nodine CF. "A Visual Concept Shapes Image Perception." Radiology 1983;146:363–368.

Kundel HL, Nodine CF, Carmody DP. "Visual Scanning, Pattern Recognition and Decision Making in Pulmonary Nodule Detection." Investigative Radiology 1978;13:175–181.

Kundel HL, Nodine CF. "Interpreting Chess Radiographs without Visual Search." Radiology 1975;116:527–532.

Kundel HL, La Follette P. "Visual Search Patterns and Experience with Radiological Images." Radiology 1972;103:523–528.

Kundel HL, Wright DJ. "The Influence of Prior Knowledge on Visual Search Strategies During Viewing of Chest Radiographs." Radiology 1969;93:315–320.

Lesgold AM, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y. "Expertise in a Complex Skill: Diagnosing X-ray Pictures." In Chi MTH, Glaser R, Farr MJ (Eds.) *The Nature of Expertise*. Hillsdale, NJ: Lawrence Erlbaum, 1988:311–142.

Lesgold AM. "Acquiring Expertise." In Anderson JR, Kosslyn SM (Eds.) *Tutorials in Learning and Memory*. San Francisco: WH Freeman, 1984:31–60.

Lesgold AM, Feltovitch PJ, Glaser R, Wang Y. "The Acquisition of Perceptual Diagnostic Skill in Radiology." LRDC Technical Report PDS-1, University of Pittsburgh, 1 September, 1981.

Lo JY, Baker JA, Kornguth PJ, Iglehart JD, Floyd Jr CE. "Predicting Breast Cancer Invasion with Artificial Neural Networks on the Basis of Mammographic Features." Radiology 1997:203:159–163.

Groner R, Groner M, Bischof WF. *Methods of Heuristics*. Hillsdale, NJ: LEA, 1983.

Haagensen CD. *Diseases of the Breast*, 2nd Edition-Revised reprint. Philadelphia: Saunders, 1986:380–382.

Hayes CC. "A Study of Solution Quality in Human Expert and Knowledge-Based System Reasoning." In Feltovich P, Ford KM, Hoffman RR (Eds.) *Expertise in Context*. Cambridge, MA: AAAI Press/The MIT Press, 1997:339–362.

Haykin S. "Neural Networks: A Comprehensive Foundation." Englewood Cliffs, NJ: Macmillan College Publishing Company, Inc., 1994.

Krupinski EA, Nodine CF, Kundel HL. "Enhancing Recognition of Lesions in Radiographic Images Using Perceptual Feedback." Optical Engineering 1998;37:813–818.

Krupinski EA. "Visual Scanning Patterns of Radiologists Searching Mammograms." Acad Radiol 1996;3:137–144.

Kundel HL, Nodine CF, Krupinski EA. "Computer-Displayed Eye Position as a Visual Aid to Pulmonary Nodule Interpretation." Invest Radiol 1990;25:890–896.

Kundel HL, Nodine CF, Thickman D, Toto L. "Searching for Lung Nodules: A Comparison of Human Performance with Random and Systematic Scanning Models." Invest Radiol 1987;22:417–422.

Kundel HL, Nodine CF. "A Visual Concept Shapes Image Perception." Radiology 1983;146:363–368.

Kundel HL, Nodine CF, Carmody DP. "Visual Scanning, Pattern Recognition and Decision Making in Pulmonary Nodule Detection." Investigative Radiology 1978;13:175–181.

Kundel HL, Nodine CF. "Interpreting Chess Radiographs without Visual Search." Radiology 1975;116:527–532.

Kundel HL, La Follette P. "Visual Search Patterns and Experience with Radiological Images." Radiology 1972;103:523–528.

Kundel HL, Wright DJ. "The Influence of Prior Knowledge on Visual Search Strategies During Viewing of Chest Radiographs." Radiology 1969;93:315–320.

Lesgold AM, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y. "Expertise in a Complex Skill: Diagnosing X-ray Pictures." In Chi MTH, Glaser R, Farr MJ (Eds.) *The Nature of Expertise*. Hillsdale, NJ: Lawrence Erlbaum, 1988:311–142.

Lesgold AM. "Acquiring Expertise." In Anderson JR, Kosslyn SM (Eds.) *Tutorials in Learning and Memory*. San Francisco: WH Freeman, 1984:31–60.

Lesgold AM, Feltovitch PJ, Glaser R, Wang Y. "The Acquisition of Perceptual Diagnostic Skill in Radiology." LRDC Technical Report PDS-1, University of Pittsburgh, 1 September, 1981.

Lo JY, Baker JA, Kornguth PJ, Iglehart JD, Floyd Jr CE. "Predicting Breast Cancer Invasion with Artificial Neural Networks on the Basis of Mammographic Features." Radiology 1997:203:159–163.

Patel VL, Groen GJ. "The General and Specific Nature of Medical Expertise: A Critical Look." In Ericsson KA, Smith J (Eds.) *Toward a General Theory of Expertise*. Cambridge: Cambridge University Press, 1991:93–125.

Parasuraman R. "Effects of Practice on Detection of Abnormalities in Chest X-Rays." Proceedings Human Factors Society 1986;309–311.

Partridge D. "Representation of Knowledge." In Boden MA (Ed.) *Artificial Intelligence*. San Diego, CA: Academic Press, 1996:55–87.

Posner MI. *Chronometric Explorations of Mind*. New York: Oxford, 1986.

Proctor RW, Dutta A. *Skill Acquisition and Human Performance*. Thousand Oaks, CA: Sage, 1995:248.

Raufaste E, Eyrolle H. "Expertise et Diagnostic Radiologique: I. Avancees Theoriques." J Radiol 1998;79:227–234.

Raufaste E, Eyrolle H. "Expertise et Diagnostic Radiologique: II. Etude Empirique." J Radiol 1998;79:235–240.

Raufaste E, Eyrolle H, Marine C. "Pertinence Generation in Radiological Diagnosis: Spreading Activation and the Nature of Expertise." Cognitive Science 1998;22:517–546.

Russell S. "Machine Learning." In Boden MA (Ed.) *Artificial Intelligence*. San Diego, CA: Academic Press, 1996:89–133.

Selfridge OG. "Pandemonium: A Paradigm for Learning." In *The Mechanisation of Thought Processes*. London: HM Stationery Office, 1959.

Smoker WRK, Berbaum KS, Luebke NH, Jacoby CG. "Spatial Perception Testing in Diagnostic Radiology." AJR 1984;143:1105–1109.

Sowden P, Davies I, Roling P. "Perceptual Learning of the Detection of Features in X-Ray Images: A Functional Role for Improvements in Adults' Visual Sensitivity?" J Experimental Psychology: Human Perception and Performance 1999 (in press).

Sternberg RJ. "Costs of expertise." In Ericsson KA (Ed.) *The Road to Excellence*. Mahwah, NJ: Lawrence Erlbaum Associates, 1996:347–354.

Tickle AB, Andrews R, Golia M, Diederick J. "The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks." IEEE Transactions on Neural Networks 1998;9:1057–1068.

Ullman S. *High-Level Vision: Object Recognition and Visual Cognition*. Cambridge, MA: MIT Press, 1996:161.

Wood BP. "Visual Expertise." Radiology 1999;211:1–3.

Zheng Y, Greenleaf JF, Gisvold JJ. "Reduction of Breast Biopsies with a Modified Self-Organizing Map." IEEE Transactions on Neural Networks 1997;8:1386–1396.